



US006604118B2

(12) **United States Patent**  
**Kleiman et al.**

(10) **Patent No.:** **US 6,604,118 B2**  
**(45) Date of Patent:** **Aug. 5, 2003**

(54) **FILE SYSTEM IMAGE TRANSFER**

(75) **Inventors:** **Steven Kleiman**, Los Altos; **David Hitz**, Portola Valley; **Guy Harris**, Mountain View; **Sean O'Malley**, Santa Clara, all of CA (US)

(73) **Assignee:** **Network Appliance, Inc.**, Sunnyvale, CA (US)

(\*) **Notice:** Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) **Appl. No.:** **09/127,497**

(22) **Filed:** **Jul. 31, 1998**

(65) **Prior Publication Data**

US 2001/0044807 A1 Nov. 22, 2001

(51) **Int. Cl.** **G06F 12/00; G06F 17/30**

(52) **U.S. Cl.** **707/203; 707/204**

(58) **Field of Search** **707/203, 204, 707/205, 100, 103, 104, 200, 201, 10, 101, 102, 104.1, 202**

(56) **References Cited**

**U.S. PATENT DOCUMENTS**

3,813,529 A	5/1974	Bartlett
3,893,024 A	7/1975	Reins et al.
4,075,691 A	2/1978	Davis et al.
4,075,704 A	2/1978	O'Leary
4,156,907 A	5/1979	Rawlings
4,333,144 A	6/1982	Whiteside
4,351,023 A	9/1982	Richer
4,377,843 A	3/1983	Garringer
4,399,503 A	8/1983	Hawley
4,456,957 A	6/1984	Schieltz
4,459,664 A	7/1984	Pottier

(List continued on next page.)

**FOREIGN PATENT DOCUMENTS**

EP 0308506 A1 3/1987

EP	0321723	6/1989
EP	0359384 A2	6/1989
EP	0359384 B1	3/1990
EP	0359384 A3	3/1990
EP	0410630 A3	1/1991
EP	0453193 A3	10/1991
EP	0453193 A2	10/1991
EP	0462917 A2	12/1991
EP	0462917 A3	12/1991

(List continued on next page.)

**OTHER PUBLICATIONS**

AT & T—UNIX System V Release 4 Programmer's Guide: Streams No Date.

AT & T Bell Laboratories Technical Journal. vol. 63 No. 08. Part 02. Dated: Oct. 1984.

Bach, Maurice J. et al "The Design of the UNIX Operating System" 1986.

Beach, Richard J. "The Message is the Medium: Multiprocess Structuring of an Interactive Paint Program". pp.: 277-287 Jul. 1982.

(List continued on next page.)

*Primary Examiner*—Safet Metjahic

*Assistant Examiner*—Uyen Le

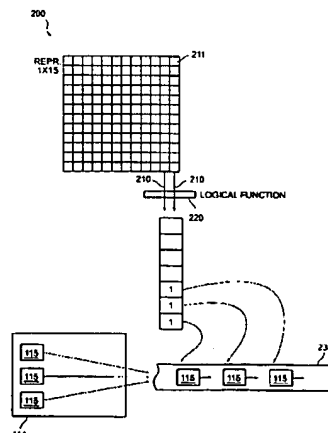
(74) *Attorney, Agent, or Firm*—Swernofsky Law Group PC

(57)

**ABSTRACT**

The invention provides a method and system for duplicating all or part of a file system while maintaining consistent copies of the file system. The file server maintains a set of snapshots, each indicating a set of storage blocks making up a consistent copy of the file system as it was at a known time. Each snapshot can be used for a purpose other than maintaining the coherency of the file system, such as duplicating or transferring a backup copy of the file system to a destination storage medium. In a preferred embodiment, the snapshots can be manipulated to identify sets of storage blocks in the file system for incremental backup or copying, or to provide a file system backup that is both complete and relatively inexpensive.

**12 Claims, 4 Drawing Sheets**



*reuse storage blocks  
but not in claims*

17

At a step 372, the file server 110 reads the incremental mirror data structure into its file system and uses that data structure to update the base storage image 220.

At a step 373, the file server 110 remounts the file system that was updated using the incremental mirror data structure.

At a flow point 380, the file server 110 is ready to continue operations with the file system restored from the base storage image 220 and the incremental mirror data structure. Volume Replication

At a flow point 350, the file server 110 is ready to perform a volume replication operation.

At a step 351, the destination file system 120 initiates the volume replication operation. The destination file system 120 sends an indicator of its newest master snapshot 210 to the file server 110, and requests the file server 110 to perform the volume replication operation.

At a step 352, the file server 110 determines if it needs to perform a volume replication operation to synchronize with a second file server 110. In this case, the second file server 110 takes the role of the destination file system 120, and initiates the volume replication operation with regard to the first file server 110.

At a step 353, the file server 110 determines its newest master snapshot 210, and its master snapshot 210 corresponding to the master snapshot 210 indicated by the destination file system 120.

If the file server 110 has at least one master snapshot 210 older than the master snapshot 210 indicated by the destination file system 120, it selects the corresponding master snapshot 210 as the newest one of those.

In this case, the method proceeds with the step 354.

If the file server 110 does not have at least one master snapshot 210 older than the master snapshot 210 indicated by the destination file system 120 (or if the destination file system 120 did not indicate any master snapshot 210), it does not select any master snapshot 210 as a corresponding master snapshot.

In this case, the method proceeds with the step 355.

At a step 354, the file server 110 performs an incremental volume copying operation, responsive to the incremental difference between the selected corresponding master snapshot 210, and the newest master snapshot 210 it has available. The method 300 proceeds with the flow point 360.

At a step 355, the file server 110 performs a full volume copying operation, responsive to the newest master snapshot 210 it has available. The method 300 proceeds with the flow point 360.

At a flow point 360, the file server 110 has completed the volume replication operation. The destination file system 120 updates its master snapshot 210 to correspond to the master snapshot 210 that was used to make the file system transfer from the file server 110.

#### Technical Appendix

A technical appendix, titled "WAFL Image Transfer," and having the inventors named as authors, forms a part of this specification, and is hereby incorporated by reference as if fully set forth herein.

#### Alternative Embodiments

Although preferred embodiments are disclosed herein, many variations are possible which remain within the concept, scope, and spirit of the invention, and these variations would become clear to those skilled in the art after perusal of this application.

What is claimed is:

1. A method for identifying storage blocks in a file system having a plurality of storage blocks, comprising the steps of:  
identifying a first storage image indicating a first set of member storage blocks selected from the plurality, the

18

first storage image including a first bit plane indicating which member storage blocks are in the first storage image;

identifying a second storage image indicating a second set of member storage blocks selected from the plurality, the second storage image including a second bit plane indicating which member storage blocks are in the second storage image;

performing a logical operation on the first bit plane and the second bit plane to determine a logical difference between the first storage image and the second storage image;

wherein at least the first storage image or the second storage image indicates member storage blocks forming a consistent file system other than an active file system.

2. A method for identifying storage blocks in a file system having a plurality of storage blocks, comprising the steps of:

identifying a first storage image indicating a first set of member storage blocks selected from the plurality, the first storage image including a first bit plane indicating which member storage blocks are in the first storage image;

identifying a second storage image indicating a second set of member storage blocks selected from the plurality, the second storage image including a second bit plane indicating which member storage blocks are in the second storage image;

performing a logical operation on the first bit plane and the second bit plane to determine a logical sum of the first storage image and the second storage image;

wherein at least the first storage image or the second storage image indicates member storage blocks forming a consistent file system other than an active file system.

3. A memory storing information including instructions, the instructions executable by a processor to identify storage blocks in a file system having a plurality of storage blocks, the instructions comprising:

identifying a first storage image indicating a first set of member storage blocks selected from the plurality, the first storage image including a first bit plane indicating which member storage blocks are in the first storage image;

identifying a second storage image indicating a second set of member storage blocks selected from the plurality, the second storage image including a second bit plane indicating which member storage blocks are in the second storage image;

performing a logical operation on the first bit plane and the second bit plane to determine a logical difference between the first storage image and the second storage image;

wherein at least the first storage image or the second storage image indicates member storage blocks forming a consistent file system other than an active file system.

4. A memory storing information including instructions, the instructions executable by a processor to identify storage blocks in a file system having a plurality of storage blocks, the instructions comprising:

identifying a first storage image indicating a first set of member storage blocks selected from the plurality, the first storage image including a first bit plane indicating which member storage blocks are in the first storage image;

19

identifying a second storage image indicating a second set of member storage blocks selected from the plurality, the second storage image including a second bit plane indicating which member storage blocks are in the second storage image; 5

performing a logical operation on the first bit plane and the second bit plane to determine a logical sum of the first storage image and the second storage image;

wherein at least the first storage image or the second storage image indicates member storage blocks forming a consistent file system other than an active file system. 10

5. An apparatus including:

a storage medium that stores a file system having a plurality of storage blocks; 15

a processor that executes instructions; and

a memory that stores the instructions, the instructions executable by the processor to identify storage blocks in the file system, the instructions comprising: (a) 20

identifying a first storage image indicating a first set of member storage blocks selected from the plurality, the first storage image including a first bit plane indicating which member storage blocks are in the first storage image, (b) identifying a second storage image indicating a second set of member storage blocks selected from the plurality, the second storage image including a second bit plane indicating which member storage blocks are in the second storage image, (c) performing a logical operation on the first bit plane and the second bit plane to determine a logical difference between the first storage image and the second storage image, wherein at least the first storage image or the second storage image indicates member storage blocks forming a consistent file system other than an active file system. 35

6. An apparatus including:

a storage medium that stores a file system having a plurality of storage blocks;

a processor that executes instructions; and 40

a memory that stores the instructions, the instructions executable by the processor to identify storage blocks

20

in the file system, the instructions comprising: (a) identifying a first storage image indicating a first set of member storage blocks selected from the plurality, the first storage image including a first bit plane indicating which member storage blocks are in the first blocks selected from the plurality, the second storage image including a second bit plane indicating which member storage blocks are in the second storage image, (c) performing a logical operation on the first bit plane and the second bit plane to determine a logical sum of the first storage image and the second storage image, wherein at least the first storage image or the second storage image indicates member storage blocks forming a consistent file system other than an active file system.

7. A method as in claim 1, wherein the first bit plane and the second bit plane each form a column of a collection of bit planes, each row in the collection representing a storage block. 20

8. A method as in claim 2, wherein the first bit plane and the second bit plane each form a column of a collection of bit planes, each row in the collection representing a storage block.

9. A memory as in claim 3, wherein the first bit plane and the second bit plane each form a column of a collection of bit planes, each row in the collection representing a storage block.

10. A memory as in claim 4, wherein the first bit plane and the second bit plane each form a column of a collection of bit planes, each row in the collection representing a storage block.

11. An apparatus as in claim 5, wherein the first bit plane and the second bit plane each form a column of a collection of bit planes, each row in the collection representing a storage block.

12. An apparatus as in claim 6, wherein the first bit plane and the second bit plane each form a column of a collection of bit planes, each row in the collection representing a storage block. 40

\* \* \* \* \*



US006289356B1

(12) **United States Patent**  
Hitz et al.

(10) Patent No.: **US 6,289,356 B1**  
(45) Date of Patent: **Sep. 11, 2001**

*This is the  
partial  
parent of  
08/854,187*

(54) **WRITE ANYWHERE FILE-SYSTEM LAYOUT**

(75) Inventors: **David Hitz**, Sunnyvale; **Michael Malcolm**, Los Altos; **James Lau**, Cupertino; **Byron Rakitzis**, Mountain View, all of CA (US)

(73) Assignee: **Network Appliance, Inc.**, Sunnyvale, CA (US)

(\*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/153,094**

(22) Filed: **Sep. 14, 1998**

**Related U.S. Application Data**

(63) Continuation of application No. 09/108,022, filed on Jun. 30, 1998, now Pat. No. 5,963,962, which is a continuation of application No. 08/454,921, filed on May 31, 1995, now Pat. No. 5,819,292, which is a continuation of application No. 08/071,643, filed on Jun. 3, 1993, now abandoned.

(51) Int. Cl.<sup>7</sup> ..... **G06F 12/00**

(52) U.S. Cl. .... **707/201; 707/202; 707/203; 707/204; 711/161; 711/162**

(58) Field of Search ..... **707/202-203, 707/201, 204; 711/163, 162, 161**

(56) **References Cited****U.S. PATENT DOCUMENTS**

4,742,447 \* 5/1988 Duvall et al. .... 709/1  
4,742,450 \* 5/1988 Duvall et al. .... 709/310

(List continued on next page.)

**FOREIGN PATENT DOCUMENTS**

0306244A2 3/1989 (EP) ..... G06F/11/00  
0308056A2 3/1989 (EP) ..... G06F/15/16

(List continued on next page.)

**OTHER PUBLICATIONS**

Borr A J: "SecureShare: safe Unix/Windows file sharing through multiprotocol locking" Proceeding of the 2nd Usenix Windows NT Symposium, proceedings of 2nd Usenix Windows NT Symposium, Seattle, WA, Aug. 3-5, 1998, pp. 117-126, XP002097387 ISBN 1-880446-95-2, 1998, Berkeley, CA, USA, Usenix Assoc. USA.

Tanner J: "CIFS: Common Internet File System" Unix Review, vol. 31, Feb. 1997, pp. 31/32, 34, XP000783952 see whole document, relevant to claim No. 1-38.

(List continued on next page.)

Primary Examiner—Jack Choules

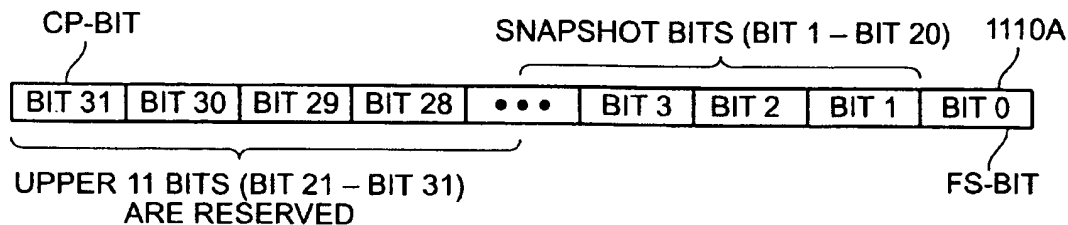
Assistant Examiner—Luke S Wassum

(74) Attorney, Agent, or Firm—Swernofsky Law Group

**(57)****ABSTRACT**

The present invention provides a method for keeping a file system in a consistent state and for creating read-only copies of a file system. Changes to the file system are tightly controlled. The file system progresses from one self-consistent state to another self-consistent state. The set of self-consistent blocks on disk that is rooted by the root inode is referred to as a consistency point. To implement consistency points, new data is written to unallocated blocks on disk. A new consistency point occurs when the fsinfo block is updated by writing a new root inode for the inode file into it. Thus, as long as the root inode is not updated, the state of the file system represented on disk does not change. The present invention also creates snapshots that are read-only copies of the file system. A snapshot uses no disk space when it is initially created. It is designed so that many different snapshots can be created for the same file system. Unlike prior art file systems that create a clone by duplicating the entire inode file and all of the indirect blocks, the present invention duplicates only the inode that describes the inode file. A multi-bit free-block map file is used to prevent data from being overwritten on disk.

**8 Claims, 40 Drawing Sheets**



*draw 5, 7*  
*block reusability*

23

In step 760, entries 2326A-2326L in block 2326 of the blkmap file 2344 are updated as illustrated in FIG. 21E. The diagram illustrates that the snapshot 2 bit (BIT2) is updated as well as the FS-BIT and CP-BIT for each entry 2326A-2326L. Thus, blocks 2304, 2308-2312, 2316-2318, 2322, and 2326 are contained in snapshot 2 whereas blocks 2306, 2314, 2320, and 2324 are not. In step 770, the dirty blocks 2308 and 2326 are written to disk.

Further processing of snapshot 2 is identical to that for generation of a consistency point illustrated in FIG. 5. In step 540, the two fsinfo blocks are flushed to disk. Thus, FIG. 21F represents the WAFL file system in a consistent state after this step. Files 2340, 2342, 2344, and 2346 of the consistent file system, after step 540 is completed, are indicated within dotted lines in FIG. 21F. In step 550, the consistency point is completed by processing inodes that were not in the consistency point.

#### Access Time Overwrites

Unix file systems must maintain an "access time" (atime) in each inode. Atime indicates the last time that the file was read. It is updated every time the file is accessed. Consequently, when a file is read the block that contains the inode in the inode file is rewritten to update the inode. This could be disadvantageous for creating snapshots because, as a consequence, reading a file could potentially use up disk space. Further, reading all the files in the file system could cause the entire inode file to be duplicated. The present invention solves this problem.

Because of atime, a read could potentially consume disk space since modifying an inode causes a new block for the inode file to be written on disk. Further, a read operation could potentially fail if a file system is full which is an abnormal condition for a file system to have occur.

In general, data on disk is not overwritten in the WAFL file system so as to protect data stored on disk. The only exception to this rule is atime overwrites for an inode as illustrated in FIGS. 23A-23B. When an "atime overwrites" occurs, the only data that is modified in a block of the inode file is the atime of one or more of the inodes it contains and the block is rewritten in the same location. This is the only exception in the WAFL system; otherwise new data is always written to new disk locations.

In FIG. 23A, the atimes 2423 and 2433 of an inode 2422 in an old WAFL inode file block 2420 and the snapshot inode 2432 that references block 2420 are illustrated. Inode 2422 of block 2420 references direct block 2410. The atime 2423 of inode 2422 is "April 30 9:15 PM" whereas the atime 2433 of snapshot inode 2432 is "May 1 10:00 AM". FIG. 23A illustrates the file system before direct buffer 2410 is accessed.

FIG. 23B illustrates the inode 2422 of direct block 2410 after direct block 2410 has been accessed. As shown in the diagram, the access time 2423 of inode 2422 is overwritten with the access time 2433 of snapshot 2432 that references it. Thus, the access time 2423 of inode 2422 for direct block 2410 is "May 1 11:23 AM".

Allowing inode file blocks to be overwritten with new atimes produces a slight inconsistency in the snapshot. The atime of a file in a snapshot can actually be later than the time that the snapshot was created. In order to prevent users from detecting this inconsistency, WAFL adjusts the atime of all files in a snapshot to the time when the snapshot was actually created instead of the time a file was last accessed. This snapshot time is stored in the inode that describes the snapshot as a whole. Thus, when accessed via the snapshot,

24

the access time 2423 for inode 2422 is always reported as "May 1 10:00 AM". This occurs both before the update when it may be expected to be "April 30 9:15 PM", and after the update when it may be expected to be "May 1 11:23 AM". When accessed through the active file system, the times are reported as "April 30 9:15 PM" and "May 1 11:23 AM" before and after the update, respectively.

In this manner, a method is disclosed for maintaining a file system in a consistent state and for creating read-only copies of the file system.

What is claimed is:

1. A method for recording a plurality of data about a plurality of blocks of data stored in storage means comprising the steps of:

maintaining a means for recording multiple usage bits per block of said storage means, wherein one bit of said multiple bits per block for each of said blocks indicates a block's membership in an active file system and one or more bits indicated membership in one or more read-only copies of a file system; and

storing, in said means for recording multiple usage bits per block, multiple bits for each of said plurality of said blocks of said storage means.

2. A method for recording a plurality of data about a plurality of blocks of data stored in storage means comprising the steps of:

maintaining a means for recording multiple usage bits per block of said storage means, wherein one bit of said multiple bits per block for each of said blocks indicates a block's membership in an active file system and one or more bits indicated membership in one or more read-only copies of a file system; and

storing, in said means for recording multiple usage bits per block, multiple bits for each of said plurality of said blocks of said storage means, at least one of said multiple bits being indicative of block reusability.

3. An apparatus having at least one processor and at least one memory coupled to said at least one processor for recording a plurality of data about a plurality of blocks of data stored in storage means, said apparatus includes:

a recording mechanism configured to record multiple usage bits per block of said storage means, responsive to said plurality of data about said plurality of said blocks of said storage means, at least one of said multiple bits being indicative of block reusability.

4. An apparatus having at least one processor and at least one memory coupled to said at least one processor for recording a plurality of data about a plurality of blocks of data stored in storage means, said apparatus includes:

a recording mechanism configured to record multiple usage bits per block of said storage means, responsive to said plurality of data about said plurality of said blocks of said storage means, wherein one bit of said multiple bits per block for each of said blocks indicates a block's membership in an active file system and one or more bits indicates membership in one or more read-only copies of a file system.

5. A computer program product including:

a computer usable storage medium having computer readable code embodied therein for causing a computer to record a plurality of data about a plurality of blocks of data stored in storage means, said computer readable code includes:

computer readable program code configured to cause said computer to effect a recording mechanism to record multiple usage bits per block of said storage means,

25

responsive to said plurality of data about said plurality of said blocks of said storage means, at least one of said multiple bits being indicative of block reusability.

6. An. A computer program product including:

a computer usable storage medium having computer readable code embodied therein for causing a computer to record a plurality of data about a plurality of blocks of data stored in storage means, said computer readable code includes:

computer readable program code configured to cause said computer to effect a recording mechanism to record multiple usage bits per block of said storage means, responsive to said plurality of data about said plurality of said blocks of said storage means, wherein one bit of said multiple bits per block for each of said blocks indicates a block's membership in an active file system and one or more bits indicates membership in one or more read-only copies of a file system.

7. A computer program product including:

a computer data signal embodied in a carrier wave having computer readable code embodied therein for causing a computer to record a plurality of data about a plurality of blocks of data stored in storage means, said computer readable code includes:

26

computer readable program code configured to cause said computer to effect a recording mechanism to record multiple usage bits per block of said storage means, responsive to said plurality of data about said plurality of said blocks of said storage means, at least one of said multiple bits being indicative of block reusability.

8. A computer program product including:

a computer data signal embodied in a carrier wave having computer readable code embodied therein for causing a computer to record a plurality of data about a plurality of blocks of data stored in storage means, said computer readable code includes:

computer readable program code configured to cause said computer to effect a recording mechanism to record multiple usage bits per block of said storage means, responsive to said plurality of data about said plurality of said blocks of said storage means, wherein one bit of said multiple bits per block for each of said blocks indicates a block's membership in an active file system and one or more bits indicates membership in one or more read-only copies of a file system.

\* \* \* \* \*



US006721764B2

**(12) United States Patent**  
**Hitz et al.****(10) Patent No.: US 6,721,764 B2****(45) Date of Patent: \*Apr. 13, 2004****(54) COPY ON WRITE FILE SYSTEM  
CONSISTENCY AND BLOCK USAGE****(75) Inventors:** David Hitz, Los Altos, CA (US);  
Michael Malcolm, Los Altos, CA (US);  
James Lau, Los Altos Hills, CA (US);  
Byron Rakitzis, Burlingame, CA (US)**(73) Assignee:** Network Appliance, Inc., Sunnyvale,  
CA (US)**(\*) Notice:** Subject to any disclaimer, the term of this  
patent is extended or adjusted under 35  
U.S.C. 154(b) by 0 days.This patent is subject to a terminal dis-  
claimer.**(21) Appl. No.: 09/954,522****(22) Filed: Sep. 11, 2001****(65) Prior Publication Data**

US 2002/0091670 A1 Jul. 11, 2002

**Related U.S. Application Data****(63)** Continuation of application No. 09/153,094, filed on Sep.  
14, 1998, now Pat. No. 6,289,356, which is a continuation  
of application No. 09/108,022, filed on Jun. 30, 1998, now  
Pat. No. 5,963,962, which is a continuation of application  
No. 08/454,921, filed on May 31, 1995, now Pat. No.  
5,819,292, which is a continuation of application No.  
08/071,643, filed on Jun. 3, 1993, now abandoned.**(51) Int. Cl.<sup>7</sup> ..... G06F 12/00; G06F 17/30****(52) U.S. Cl. .... 707/202; 707/203; 707/204;**  
714/15; 714/20**(58) Field of Search .... 707/203, 202;**  
714/15-20**(56) References Cited****U.S. PATENT DOCUMENTS**4,814,971 A \* 3/1989 Thatte ..... 714/15  
4,878,167 A \* 10/1989 Kapulka et al. .... 714/165,008,786 A \* 4/1991 Thatte ..... 711/162  
5,043,871 A \* 8/1991 Nishigaki et al. .... 707/202  
5,043,876 A \* 8/1991 Terry ..... 707/201  
5,163,148 A \* 11/1992 Walls ..... 707/204  
5,182,805 A \* 1/1993 Campbell ..... 711/163  
5,218,695 A \* 6/1993 Noveck et al. .... 707/205  
5,369,757 A \* 11/1994 Spiro et al. .... 714/19

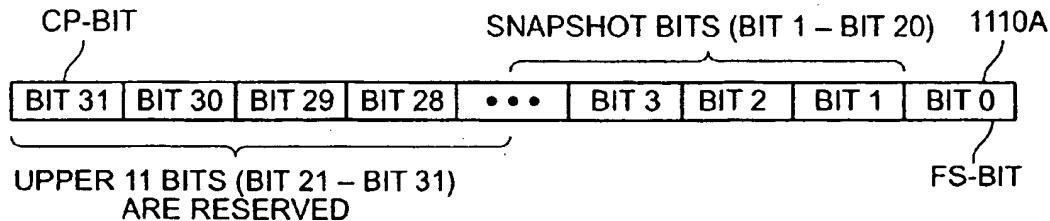
(List continued on next page.)

**OTHER PUBLICATIONS**IBM technical Disclosure Bulletin, "Microkernel Extension  
for Cloning", vol. 37, No. 4B, Apr. 1994, pp. 509-512.\*  
Hitz, D. et al. "File System Design for an NFS File Server  
Appliance", USENIX, Jan. 19, 1994.\*

(List continued on next page.)

*Primary Examiner*—John Breene*Assistant Examiner*—Luke S Wassum**(74) Attorney, Agent, or Firm**—Swernofsky Law Group PC**(57) ABSTRACT**

The present invention provides a method for keeping a file system in a consistent state and for creating read-only copies of a file system. Changes to the file system are tightly controlled. The file system progresses from one self-consistent state to another self-consistent state. The set of self-consistent blocks on disk that is rooted by the root inode is referred to as a consistency point. To implement consistency points, new data is written to unallocated blocks on disk. A new consistency point occurs when the fsinfo block is updated by writing a new root inode for the inode file into it. Thus, as long as the root inode is not updated, the state of the file system represented on disk does not change. The present invention also creates snapshots that are read-only copies of the file system. A snapshot uses no disk space when it is initially created. It is designed so that many different snapshots can be created for the same file system. Unlike prior art file systems that create a clone by duplicating the entire inode file and all of the indirect blocks, the present invention duplicates only the inode that describes the inode file. A multi-bit free-block map file is used to prevent data from being overwritten on disk.

**22 Claims, 40 Drawing Sheets**cl 2 block reusability  
#9  
#16

23

space. Further, reading all the files in the file system could cause the entire inode file to be duplicated. The present invention solves this problem.

Because of a time, a read could potentially consume disk space since modifying an inode causes a new block for the inode file to be written on disk. Further, a read operation could potentially fail if a file system is full which is an abnormal condition for a file system to have occur.

In general, data on disk is not overwritten in the WAFL file system so as to protect data stored on disk. The only exception to this rule is a time overwrites for an inode as illustrated in FIGS. 23A–238. When an “a time overwrites” occurs, the only data that is modified in a block of the inode file is the a time of one or more of the inodes it contains and the block is rewritten in the same location. This is the only exception in the WAFL system; otherwise new data is always written to new disk locations.

In FIG. 23A, the times 2423 and 2433 of an inode 2422 in an old WAFL inode file block 2420 and the snapshot inode 2432 that references block 2420 are illustrated. Inode 2422 of block 2420 references direct block 2410. The a time 2423 of inode 2422 is “4/30 9:15 PM” whereas the a time 2433 of snapshot inode 2432 is “5/1 10:00 AM”. FIG. 23A illustrates the file system before direct buffer 2410 is accessed.

FIG. 23B illustrates the inode 2422 of direct block 2410 after direct block 2410 has been accessed. As shown in the diagram, the access time 2423 of inode 2422 is overwritten with the access time 2433 of snapshot 2432 that references it. Thus, the access time 2423 of inode 2422 for direct block 2410 is “5/1 11:23 AM”.

Allowing inode file blocks to be overwritten with new a times produces a slight inconsistency in the snapshot. The a time of a file in a snapshot can actually be later than the time that the snapshot was created. In order to prevent users from detecting this inconsistency, WAFL adjusts the a time of all files in a snapshot to the time when the snapshot was actually created instead of the time a file was last accessed. This snapshot time is stored in the inode that describes the snapshot as a whole. Thus, when accessed via the snapshot, the access time 2423 for inode 2422 is always reported as “5/1 10:00AM”. This occurs both before the update when it may be expected to be “4/30 9:15PM”, and after the update when it may be expected to be “5/1 11:23AM”. When accessed through the active file system, the times are reported as “4/30 9:15PM” and “5/1 11:23AM” before and after the update, respectively.

In this manner, a method is disclosed for maintaining a file system in a consistent state and for creating read-only copies of the file system.

What is claimed is:

1. A method for recording a plurality of data about a plurality of blocks of data stored in a storage system, comprising the step of maintaining multiple usage bits for each of said plurality of blocks;

wherein one bit of said multiple usage bits for each of said plurality of blocks indicates a block's membership in an active file system and plural bits of said multiple usage bits for each of said plurality of blocks indicate membership in plural read-only copies of a file system; and

wherein said multiple usage bits for each of said plurality of blocks are maintained in said active file system.

2. A method as in claim 1, wherein one or more bits of said multiple usage bits for each of said plurality of blocks further indicate block reusability.

3. A method for generating a consistency point for a storage system, comprising the steps of:

marking a plurality of inodes pointing to a plurality of modified blocks in a file system stored on said storage system as being in a consistency point;

24

flushing regular files to said storage system;

flushing special files to said storage system;

flushing at least one block of file system information to said storage system; and

queuing dirty inodes after said step of marking and before said step of flushing at least one block of file system information;

wherein said step of flushing said special files to said storage system further comprises the steps of:

pre-flushing an inode for a blockmap file to an inode file;

allocating space on said storage system for all dirty blocks in said inode and said blockmap files;

flushing said inode for said blockmap file again

updating a plurality of entries in said blockmap file wherein each entry of said plurality of entries represents a block in said storage system; and

writing all dirty blocks in said blockmap file and said inode file to said storage system.

4. A method of maintaining data in a storage system, comprising the steps of:

maintaining a root inode and inodes for a file system in the storage system, the root inode pointing directly or indirectly to the inodes, and each inode storing file data, pointing to one or more blocks in the storage system that store file data, or pointing to other inodes, with different levels of indirection possible for the inodes to point to the blocks; and

maintaining an inode map and a block map for the file; wherein when said inodes are maintained in the storage system, each inode only points to blocks in the storage system having a same level of indirection as other blocks pointed to by that inode.

5. A method as in claim 4, further comprising the step of creating a snapshot of the file system by copying the root inode.

6. A method as in claim 5, wherein the block map indicates membership of blocks in one or more snapshots.

7. A method as in claim 5, further comprising the step of deleting a snapshot from the storage system, wherein blocks that are only part of the deleted snapshot are released for re-use by the storage system.

8. A memory storing information including instructions, the instructions executable by a processor to record a plurality of data about a plurality of blocks of data stored in a storage system, the instructions comprising the step of maintaining multiple usage bits for each of said plurality of blocks;

wherein one bit of said multiple usage bits for each of said plurality of blocks indicates a block's membership in an active file system and plural bits of said multiple usage bits for each of said plurality of blocks indicate membership in plural read-only copies of a file system; and

wherein said multiple usage bits for each of said plurality of blocks are maintained in said active file system.

9. A memory as in claim 8, wherein one or more bits of said multiple usage bits for each of said plurality of blocks further indicate block reusability.

10. A memory storing information including instructions, the instructions executable by a processor to generate a consistency point for a storage system, the instructions comprising the steps of:

marking a plurality of inodes pointing to a plurality of modified blocks in a file system stored on said storage system as being in a consistency point;

flushing regular files to said storage system;



25

flushing special files to said storage system;  
 flushing at least one block of file system information to said storage system; and  
 queuing dirty inodes after said step of marking and before said step of flushing at least one block of file system information;  
 wherein said step of flushing said special files to said storage system further comprises the steps of:  
 pre-flushing, an inode for a blockmap file to an inode file;  
 allocating space on said storage system for all dirty blocks in said mode and said blockmap files;  
 flushing said inode for said blockmap file again;  
 updating a plurality of entries in said blockmap file wherein each entry of said plurality of entries represents a block in said storage system; and  
 writing all dirty blocks in said blockmap file and said inode file to said storage system.

11. A memory storing information including instructions, the instructions executable by a processor to maintain data in a storage system, the instructions comprising the steps of:  
 maintaining a root inode mode and inodes for a file system in the storage system, the root inode pointing directly or indirectly to the inodes, and each inode storing file data, pointing to one or more blocks in the storage system that store file data, or pointing to other inodes, with different levels of indirection possible for the inodes to point to the blocks; and  
 maintaining an inode map and a block map for the file system;  
 wherein when said inodes are maintained in the storage system, each inode only points to blocks in the storage system having a same level of indirection as other blocks pointed to by that inode.

12. A memory as in claim 11, wherein the instructions further comprise the step of creating a snapshot or the file system by copying the root inode.

13. A memory as in claim 12, wherein the block map indicates membership of blocks in one or more snapshots.

14. A memory as in claim 12, wherein the instructions further comprise the step of deleting a snapshot from the storage system, wherein blocks that are only part of the deleted snapshot are released for re-use by the storage system.

15. A system comprising:  
 a processor;  
 a storage system; and  
 a memory storing information including instructions, the instructions executable by the processor to record a plurality of data about a plurality of blocks of data stored in the storage system, the instructions comprising step of maintaining multiple usage bits for each of said plurality of blocks;  
 wherein one bit of said multiple usage bits for each of said plurality of blocks indicates a block's membership in an active file system and plural bits of said multiple usage bits for each of said plurality of blocks indicate membership in plural read-only copies of a file system; and  
 wherein said multiple usage bits for each of said plurality of blocks are maintained in said active file system.

16. A system as in claim 15, wherein one or more bits of said multiple usage bits for each of said plurality of blocks further indicate block reusability.

17. A system comprising:  
 a processor;  
 a storage system; and

26

a memory storing information including instruction, the instructions executable by the processor to generate a consistency point for the storage system, the instructions comprising the steps of: (a) marking a plurality of inodes pointing to a plurality of modified blocks in a file system stored on said storage system as being in a consistency point; (b) flushing regular files to said storage system; (c) flushing special files to said storage system; (d) flushing at least one block of file system information to said storage system; and (e) queuing dirty inodes after said step of marking and before said step of flushing at least one block of file system information; and  
 wherein said step of flushing said special files to said storage system further comprises the steps of: (f) pre-flushing an inode for a blockmap file to an inode file; (g) allocating space on said storage system for all dirty blocks in said inode and said blockmap files; (h) flushing said inode for said blockmap file again; (i) updating a plurality of entries in said blockmap file wherein each entry of said plurality of entries represents a block in said storage system; and (j) writing all dirty blocks in said blockmap file and said inode file to said storage system.

18. A system comprising:  
 a processor;  
 a storage system; and  
 a memory storing information including instructions, the instructions executable by the processor to maintain data in the storage system, the instructions comprising the steps of: (a) maintaining a root inode and inodes for a file system in the storage system, the root inode pointing directly or indirectly to the inodes, and each inode storing file data, pointing to one or more blocks in the storage system that store file data, or pointing to other inodes, with different levels of indirection possible for the inodes to point to the blocks; and (b) maintaining an inode map and a block map for the file system;  
 wherein when said inodes are maintained in the storage system, each inode only points to blocks in the storage system having a same level of indirection as other blocks pointed to by that inode.

19. A system as in claim 18, wherein the instructions further comprise the step of creating a snapshot of the file system by copying the root inode.

20. A system as in claim 19, wherein the block map indicates membership of blocks in one or more snapshots.

21. A system as in claim 19, wherein the instructions further comprise the step of deleting a snapshot from the storage system, wherein blocks that are only part of the deleted snapshot are released for re-use by the storage system.

22. A system for maintaining data in storage means, comprising:  
 means for maintaining a root inode and inodes for a file system in the storage systems, the root inode pointing to inodes, and each inode storing file data, pointing to one or more blocks in the storage means that store file data, or pointing to other inodes; and  
 means for maintaining an inode map and a block map for the file system;  
 wherein when said inodes are maintained in the storage system, each inode only points to blocks having a same level of indirection as other blocks pointed to by that inode.

\* \* \* \* \*



US005819292A

**United States Patent** [19]

Hitz et al.

[11] **Patent Number:** 5,819,292[45] **Date of Patent:** Oct. 6, 1998

[54] **METHOD FOR MAINTAINING CONSISTENT STATES OF A FILE SYSTEM AND FOR CREATING USER-ACCESSIBLE READ-ONLY COPIES OF A FILE SYSTEM**

[75] **Inventors:** David Hitz, Sunnyvale; Michael Malcolm, Los Altos; James Lau, Cupertino; Byron Rakitzis, Mountain View, all of Calif.

[73] **Assignee:** Network Appliance, Inc., Santa Clara, Calif.

[21] **Appl. No.:** 454,921

[22] **Filed:** May 31, 1995

**Related U.S. Application Data**

[63] Continuation of Ser. No. 71,643, Jun. 3, 1993, abandoned.

[51] **Int. Cl.<sup>6</sup>** ..... G06F 17/30

[52] **U.S. Cl.** ..... 707/203; 707/205

[58] **Field of Search** ..... 395/621, 619

[56] **References Cited****U.S. PATENT DOCUMENTS**

4,075,691	2/1978	Davis et al.	395/884
4,156,907	5/1979	Rawlings et al.	395/200.01
4,333,144	6/1982	Whiteside et al.	395/672
4,377,843	3/1983	Garringer et al.	395/890
4,399,503	8/1983	Hawley	395/440
4,456,957	6/1984	Schieltz	395/824
4,459,664	7/1984	Pottier et al.	395/675
4,488,231	12/1984	Yu et al.	395/868
4,527,232	7/1985	Bechtolsheim	395/416
4,550,368	10/1985	Bechtolsheim	395/416
4,685,125	8/1987	Zave	379/96
4,710,868	12/1987	Cocke et al.	395/311
4,719,569	1/1988	Ludemann et al.	395/729
4,742,447	5/1988	Duvall et al.	364/200
4,761,785	8/1988	Clark et al.	371/51
4,766,534	8/1988	DeBenedictis	395/200.14
4,780,821	10/1988	Crossley	395/670

(List continued on next page.)

**OTHER PUBLICATIONS**

Bach, "The Design of the UNIX Operating System", Prentice-Hall, 1986, pp. 38-90 and 325-329, Jan. 1986.

Beach, Richard J., et al., "The Message is The Medium: Multiprocess Structuring of an Interactive Paint Program", Computer Graphics, vol. 16, No. 3, Jul. 1982, pp. 277-287.

Britton, Dianne E. et al., "A Interprocess Communication Facility for Distributed Applications", RCA Laboratories, David Sarnoff Research Center, Princeton, N.J., Copyright 1980 IEEE, Jan. 1980.

Carr, Robert et al., The Power of PenPoint, Ch 5, the Class Manager, Copyright 1991, Addison-Wesley, Jan. 1991.

Cashin, "Interprocess Communication", Bell-Northern Research, Document No. 8005014, Jan. 1980.

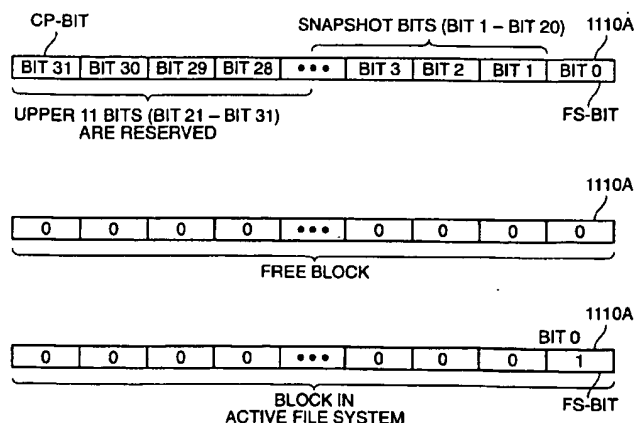
(List continued on next page.)

**Primary Examiner**—Paul R. Lintz

**Attorney, Agent, or Firm**—D'Alessandro & Ritchie

[57] **ABSTRACT**

A method is disclosed for maintaining consistent states of a file system. The file system progresses from one self-consistent state to another self-consistent state. The set of self-consistent blocks on disk that is rooted by a root inode is referred to as a consistency point. The root inode is stored in a file system information structure. To implement consistency points, new data is written to unallocated blocks on disk. A new consistency point occurs when the file system information structure is updated by writing a new root inode into it. Thus, as long as the root inode is not updated, the state of the file system represented on disk does not change. The method also creates snapshots that are user-accessible read-only copies of the file system. A snapshot uses no disk space when it is initially created. It is designed so that many different snapshots can be created for the same file system. Unlike prior art file systems that create a done by duplicating an entire inode file and all indirect blocks, the method of the present invention duplicates only the inode that describes the inode file. A multi-bit free-block map file is used to prevent data referenced by snapshots from being overwritten on disk.

**20 Claims, 39 Drawing Sheets**

In step 760, entries 2326A-2326L in block 2326 of the blkmap file 2344 are updated as illustrated in FIG. 21E. The diagram illustrates that the snapshot 2 bit (BIT2) is updated as well as the FS-BIT and CP-BIT for each entry 2326A-2326L. Thus, blocks 2304, 2308-2312, 2316-2318, 2322, and 2326 are contained in snapshot 2 whereas blocks 2306, 2314, 2320, and 2324 are not. In step 770, the dirty blocks 2308 and 2326 are written to disk, as indicated in FIG. 21F.

Further processing of snapshot 2 is identical to that for generation of a consistency point illustrated in FIG. 5. In step 540, the two fsinfo blocks are flushed to disk. Thus, FIG. 21F represents the WAFL file system in a consistent state after this step. Files 2340, 2342, 2344, and 2346 of the consistent file system, after step 540 is completed, are indicated within dotted lines in FIG. 21F. In step 550, the consistency point is completed by processing inodes that were not in the consistency point.

#### Access Time Overwrites

Unix file systems must maintain an "access time" (atime) in each inode. Atime indicates the last time that the file was read. It is updated every time the file is accessed. Consequently, when a file is read the block that contains the inode in the inode file is rewritten to update the inode. This could be disadvantageous for creating snapshots because, as a consequence, reading a file could potentially use up disk space. Further, reading all the files in the file system could cause the entire inode file to be duplicated. The present invention solves this problem.

Because of atime, a read could potentially consume disk space since modifying an inode causes a new block for the inode file to be written on disk. Further, a read operation could potentially fail if a file system is full which is an abnormal condition for a file system to have occur.

In general, data on disk is not overwritten in the WAFL file system so as to protect data stored on disk. The only exception to this rule is atime overwrites for an inode as illustrated in FIGS. 23A-23B. When an "atime overwrites" occurs, the only data that is modified in a block of the inode file is the atime of one or more of the inodes it contains and the block is rewritten in the same location. This is the only exception in the WAFL system; otherwise new data is always written to new disk locations.

In FIG. 23A, the atimes 2423 and 2433 of an inode 2422 in an old WAFL inode file block 2420 and the snapshot inode 2432 that references block 2420 are illustrated. Inode 2422 of block 2420 references direct block 2410. The atime 2423 of inode 2422 is "4/30 9:15 PM" whereas the atime 2433 of snapshot inode 2432 is "5/1 10:00 AM". FIG. 23A illustrates the file system before direct buffer 2410 is accessed.

FIG. 23B illustrates the inode 2422 of direct block 2410 after direct block 2410 has been accessed. As shown in the diagram, the access time 2423 of inode 2422 is overwritten with the access time at which direct block 2410 was accessed. Thus, the access time 2423 of inode 2422 for direct block 2410 is "5/1 11:23 AM".

Allowing inode file blocks to be overwritten with new atimes produces a slight inconsistency in the snapshot. The atime of a file in a snapshot can actually be later than the time that the snapshot was created. In order to prevent users from detecting this inconsistency, WAFL adjusts the atime of all files in a snapshot to the time when the snapshot was actually created instead of the time a file was last accessed. This snapshot time is stored in the inode that describes the snapshot as a whole. Thus, when accessed via the snapshot, the access time 2423 for inode 2422 is always reported as

"5/1 10:00AM". This occurs both before the update when it may be expected to be "4/30 9:15PM", and after the update when it may be expected to be "5/1 11:23AM". When accessed through the active file system, the times are reported as "4/30 9:15PM" and "5/1 11:23AM" before and after the update, respectively.

In this manner, a method is disclosed for maintaining a file system in a consistent state and for creating read-only copies of the file system.

We claim:

1. A method for recording a plurality of data about a plurality of blocks of data stored in storage means comprising the steps of:

maintaining a means for recording multiple usage bits per block of said storage means, wherein one bit of said multiple bits per block for each of said blocks indicates a block's membership in an active file system and one or more bits indicated membership in one or more read-only copies of a file system; and

storing in said means for recording multiple usage bits per block multiple bits for each of a plurality of said blocks of said storage means.

2. A method for maintaining a file system stored in non-volatile storage means at successive consistency points said file system comprising blocks of data, said blocks of data comprising blocks of regular file data and blocks of meta-data file data referencing said blocks of data of said file system, said meta file data comprising a file system information structure comprising data describing said file system at a first consistency point said computer system further comprising memory means, said method comprising the steps of:

maintaining a plurality of modified blocks of regular file data and meta-data file data in said memory means, said modified blocks of data comprising blocks of data modified from said first consistency point;

designating as dirty blocks of meta-data file data referencing said modified blocks of regular file data and meta-data file data, said dirty blocks of meta-data file data comprising blocks of meta-data file data to be included in a second consistency point;

copying said modified blocks of regular file data referenced by said dirty blocks of meta-data file data to free blocks of said non-volatile storage means;

copying blocks comprising said modified blocks of meta-data file data referenced by said dirty blocks of meta-data file data to free blocks of said non-volatile storage means;

modifying a copy of said file system information structure maintained in said memory means to reference said dirty blocks of meta-data file data;

copying said modified file system information structure to said non-volatile storage means.

3. The method of claim 2 wherein said blocks of meta-file data comprise one or more blocks of inode file data and one or more blocks of blockmap file data and wherein said step of copying said modified blocks of meta-data file data to free blocks of said non-volatile storage means further comprises the steps of:

copying an inode referencing one or more blocks of blockmap file data to a block of inode file data maintained in said memory means;

allocating free blocks of said non-volatile storage means for said block of inode file data and one or more modified blocks of blockmap file data;

25

updating said inode referencing said one or more blocks of blockmap file data to reference said one or more free blocks of said non-volatile storage means allocated to said one or more modified blocks of blockmap file data; copying said updated inode to said block of inode file data;

updating said one or more blocks of blockmap file data; writing said updated one or more blocks of blockmap file data and said block of inode file data to said allocated free blocks of said non-volatile storage means.

4. A method for maintaining a file system comprising blocks of data stored in blocks of a non-volatile storage means at successive consistency points comprising the steps of:

storing a first file system information structure for a first consistency point in said non-volatile storage means, said first file system information structure comprising data describing a layout of said file system at said first consistency point of said file system;

writing blocks of data of said file system that have been modified from said first consistency point as of the commencement of a second consistency point to free blocks of said non-volatile storage means;

storing in said non-volatile storage means a second file system information structure for said second consistency point, said second file system information structure comprising data describing a layout said file system at said second consistency point of said file system.

5. The method of claim 4 wherein said step of storing said first file system information structure in said non-volatile storage means comprises the step of:

storing first and second copies of said first file system information structure at first and second locations respectively of said non-volatile storage means;

and wherein said step of storing said second file system information structure in said non-volatile storage means comprises the steps of:

overwriting said first copy of said first file system information structure with a first copy of said second file system information structure; and

overwriting said second copy of said first file system information structure with a second copy of said second file system information structure.

6. The method of claim 5 wherein said first and second locations of said non-volatile storage means comprise fixed predetermined locations of said non-volatile storage means.

7. The method of claim 5 wherein each copy of said file system information structure comprises means for determining a most recent version of said file system information structure and means for determining validity of said file system information structure, further comprising the steps of:

after a system failure, reading said first and second copies of said file system information structure from said first and second locations of said non-volatile storage means;

determining a most recent valid file system information structure from said first and second copies of said file system information structure.

26

8. A method for creating a plurality of read-only copies of a file system stored in blocks of a non-volatile storage means, said file system comprising meta-data identifying blocks of said non-volatile storage means used by said file system, comprising the steps of:

storing meta-data for successive states of said file system in said non-volatile storage means;

making a copy of said meta-data at each of a plurality of said states of said file system;

for each of said copies of said meta-data at a respective state of said file system, marking said blocks of said non-volatile storage means identified in said meta-data as comprising a respective read-only copy of said file system.

9. The method of claim 8 wherein said step of marking said blocks comprising a respective read-only copy of said file system comprises placing an appropriate entry in a means for recording multiple usage bits per block of said non-volatile storage means.

10. The method of claim 9 wherein said means for recording multiple usage bits per block of said non-volatile storage means comprises a blockmap comprising multiple bit entries for each block.

11. The method of claim 8 wherein said meta-data comprises pointers to a hierarchical tree of blocks comprising said file system.

12. The method of claim 8 wherein said meta-data comprises structures representing files of said file system.

13. The method of claim 12 wherein said structures representing files of said file system comprise inodes.

14. The method of claim 8 further comprising the step of: preventing overwriting of said blocks marked as belonging to a read-only copy of said file system.

15. The method of claim 8 comprising the step of unmarking said blocks marked as belonging to a read only copy of said file system when said read only copy of said file system is no longer needed.

16. The method of claim 8 wherein a plurality of said blocks marked as belonging to a read-only copy of said file system comprise data ancillary to said file system, said method further including the steps of:

allowing said ancillary data to be overwritten; and

otherwise preventing overwriting of said blocks marked as comprising a read only copy of said file system.

17. The method of claim 16 wherein said ancillary data comprises access time data.

18. The method of claim 8 wherein said meta-data comprises a root structure referencing structures representing files of said file system, and wherein said copies of said meta-data comprise copies of said root structure.

19. The method of claim 18 wherein said root structure comprises a root inode.

20. The method of claim 8 further comprising the step of using one or more of said read-only copies of said file system to back-up said blocks comprising one or more consistency points of said file system.

\* \* \* \* \*



US005963962A

**United States Patent** [19]**Hitz et al.**[11] **Patent Number:** **5,963,962**[45] **Date of Patent:** **\*Oct. 5, 1999**[54] **WRITE ANYWHERE FILE-SYSTEM LAYOUT**WO 9313475 8/1993 WIPO .  
WO 9821656 5/1998 WIPO .[75] **Inventors:** **David Hitz**, Sunnyvale; **Michael Malcolm**, Los Altos; **James Lau**, Cupertino; **Byron Rakitzis**, Mountain View, all of Calif.[73] **Assignee:** **Network Appliance, Inc.**, Santa Clara, Calif.[\*] **Notice:** This patent is subject to a terminal disclaimer.[21] **Appl. No.:** **09/108,022**[22] **Filed:** **Jun. 30, 1998****Related U.S. Application Data**

[63] Continuation of application No. 08/454,921, May 31, 1995, Pat. No. 5,819,292.

[51] **Int. Cl.<sup>6</sup>** ..... **G06F 17/30**[52] **U.S. Cl.** ..... **707/202; 707/201; 707/204; 395/182.16**[58] **Field of Search** ..... **707/202, 201, 707/204; 345/182.16**[56] **References Cited****U.S. PATENT DOCUMENTS**5,146,588 9/1992 Crater et al. .  
5,335,235 8/1994 Arnott .  
5,642,501 6/1997 Doshi et al. .... 707/8**FOREIGN PATENT DOCUMENTS**359384 3/1990 European Pat. Off. .  
453193 10/1991 European Pat. Off. .  
552580 7/1993 European Pat. Off. .  
629956 A2 12/1994 European Pat. Off. .  
629956 A3 12/1994 European Pat. Off. .  
747829 12/1996 European Pat. Off. .  
756235 1/1997 European Pat. Off. .  
WO 9113404 5/1991 WIPO .**OTHER PUBLICATIONS**

B. Srinivasan, et al.: "Recoverable file system for micro-processor systems" Microprocessors and Microsystems., vol. 9, No. 4, May 1985, London, GB, pp. 179-183, XP002031805 see p. 181, left-hand col., line 12 -p. 182, right-hand col., line 30.

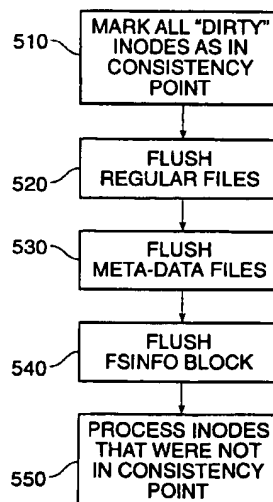
R. Nass: Connect Disk Arrays to Eisa or PCI Buses Electronic Design, vol. 41, No. 23, Nov. 1993, Cleveland, OH, USA pp. 152-154, XP000417908 see p. 152 -p. 153, right-hand col., line 10.

*Primary Examiner*—Paul R. Lintz*Attorney, Agent, or Firm*—Swernofsky Law Group

## [57]

**ABSTRACT**

The present invention provides a method for keeping a file system in a consistent state and for creating read-only copies of a file system. Changes to the file system are tightly controlled. The file system progresses from one self-consistent state to another self-consistent state. The set of self-consistent blocks on disk that is rooted by the root inode is referred to as a consistency point. To implement consistency points, new data is written to unallocated blocks on disk. A new consistency point occurs when the fsinfo block is updated by writing a new root inode for the inode file into it. Thus, as long as the root inode is not updated, the state of the file system represented on disk does not change. The present invention also creates snapshots that are read-only copies of the file system. A snapshot uses no disk space when it is initially created. It is designed so that many different snapshots can be created for the same file system. Unlike prior art file systems that create a clone by duplicating the entire inode file and all of the indirect blocks, the present invention duplicates only the inode that describes the inode file. A multi-bit free-block map file is used to prevent data from being overwritten on disk.

**2 Claims, 39 Drawing Sheets**

*reuse block  
but not in claim*

## 23

In general, data on disk is not overwritten in the WAFL file system so as to protect data stored on disk. The only exception to this rule is atime overwrites for an inode as illustrated in FIGS. 23A–23B. When an “atime overwrites” occurs, the only data that is modified in a block of the inode file is the atime of one or more of the inodes it contains and the block is rewritten in the same location. This is the only exception in the WAFL system; otherwise new data is always written to new disk locations.

In FIG. 23A, the atimes 2423 and 2433 of an inode 2422 in an old WAFL inode file block 2420 and the snapshot inode 2432 that references block 2420 are illustrated. Inode 2422 of block 2420 references direct block 2410. The atime 2423 of inode 2422 is “4/30 9:15 PM” whereas the atime 2433 of snapshot inode 2432 is “5/1 10:00 AM”. FIG. 23A illustrates the file system before direct buffer 2410 is accessed.

FIG. 23B illustrates the inode 2422 of direct block 2410 after direct block 2410 has been accessed. As shown in the diagram, the access time 2423 of inode 2422 is overwritten with the access time 2433 of snapshot 2432 that references it. Thus, the access time 2423 of inode 2422 for direct block 2410 is “5/1 11:23 AM”.

Allowing inode file blocks to be overwritten with new atimes produces a slight inconsistency in the snapshot. The time of a file in a snapshot can actually be later than the time that the snapshot was created. In order to prevent users from detecting this inconsistency, WAFL adjusts the atime of all files in a snapshot to the time when the snapshot was actually created instead of the time a file was last accessed. This snapshot time is stored in the inode that describes the snapshot as a whole. Thus, when accessed via the snapshot, the access time 2423 for inode 2422 is always reported as “5/1 10:00AM”. This occurs both before the update when it may be expected to be “4/30 9:15PM”, and after the update when it may be expected to be “5/1 11:23AM”. When

## 24

accessed through the active file system, the times are reported as “4/30 9:15PM” and “5/1 11:23AM” before and after the update, respectively.

In this manner, a method is disclosed for maintaining a file system in a consistent state and for creating read-only copies of the file system.

We claim:

1. A method for generating a consistency point comprising the steps of:

marking a plurality of inodes pointing to a plurality of modified blocks in a file system as being in a consistency point;

flushing regular files to storage means;

flushing special files to said storage means;

flushing at least one block of file system information to said storage means; and,

requeueing any dirty inodes that were not part of said consistency point.

2. The method of claim 1 wherein said step of flushing said special files to said storage means further comprises the steps of:

pre-flushing an inode for a blockmap file to an inode file; allocating space on said storage means for all dirty blocks in said inode and said blockmap files;

flushing said inode for said blockmap file again;

updating a plurality of entries in said blockmap file wherein each entry of said plurality of entries represents a block on said storage means; and,

writing all dirty blocks in said blockmap file and said inode file to said storage means.

\* \* \* \* \*

## [54] IMPROVED RECOVERABLE VIRTUAL MEMORY HAVING PERSISTENT OBJECTS

[75] Inventor: Satish M. Thatte, Richardson, Tex.

[73] Assignee: Texas Instruments Incorporated, Dallas, Tex.

[21] Appl. No.: 932,813

[22] Filed: Nov. 19, 1986

## Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 774,828, Sep. 11, 1985.

[51] Int. Cl.<sup>3</sup> ..... G06F 11/00[52] U.S. Cl. .... 364/200; 364/365.3;  
364/256.3

[58] Field of Search ... 364/200 MS File, 900 MS File

## [56] References Cited

## U.S. PATENT DOCUMENTS

4,121,286	10/1978	Venton et al.	364/200
4,376,297	3/1983	Anderson et al.	364/200
4,459,658	7/1984	Gabbe et al.	364/200
4,674,038	6/1987	Brelsford et al.	364/200
4,703,481	10/1987	Fremont	364/200
4,814,971	3/1989	Thatte	364/200

## OTHER PUBLICATIONS

Terrence W. Pratt, Programming Languages Design and Implementation, 1975, p. 112.

Raymond A. Lorie, ACM Transactions on Database Systems, vol. 2, No. 1, Mar. 1977, pp. 91-104.

Joost Verhofstad, ACM Computing Surveys, vol. 10, No. 2, Jun. 1978, pp. 165-195.

Jim Gray, ACM Computing Surveys, vol. 13, No. 2, Jun. 1981, pp. 223-242.

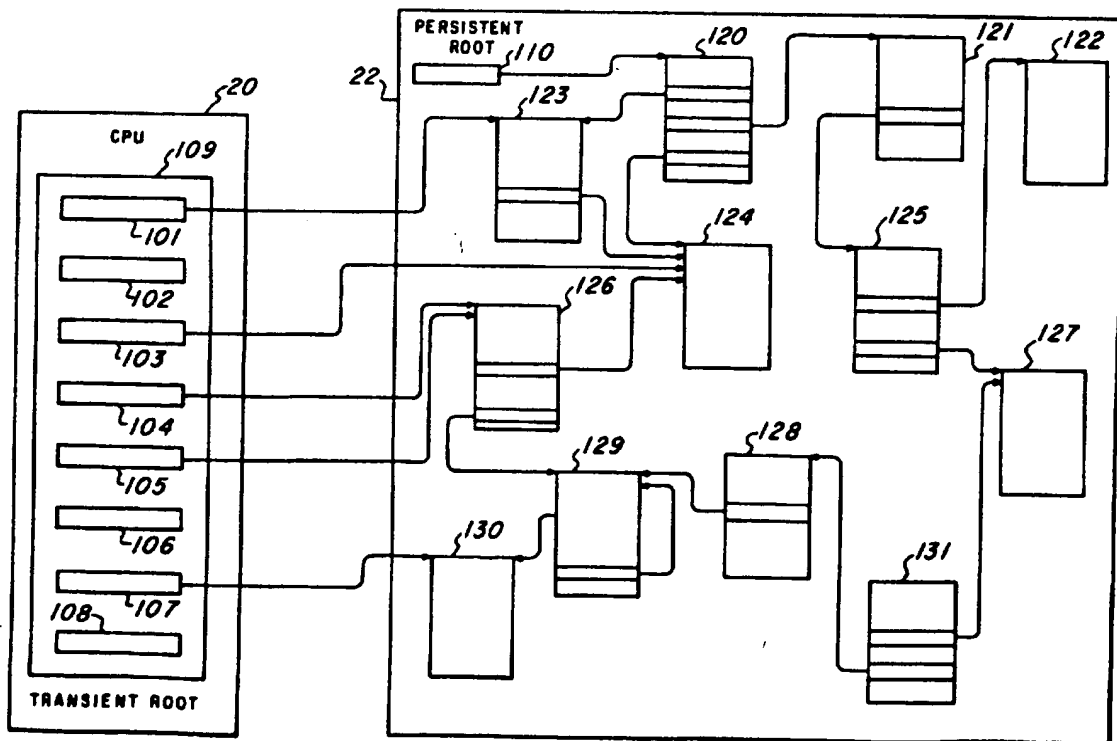
Primary Examiner—David Y. Eng

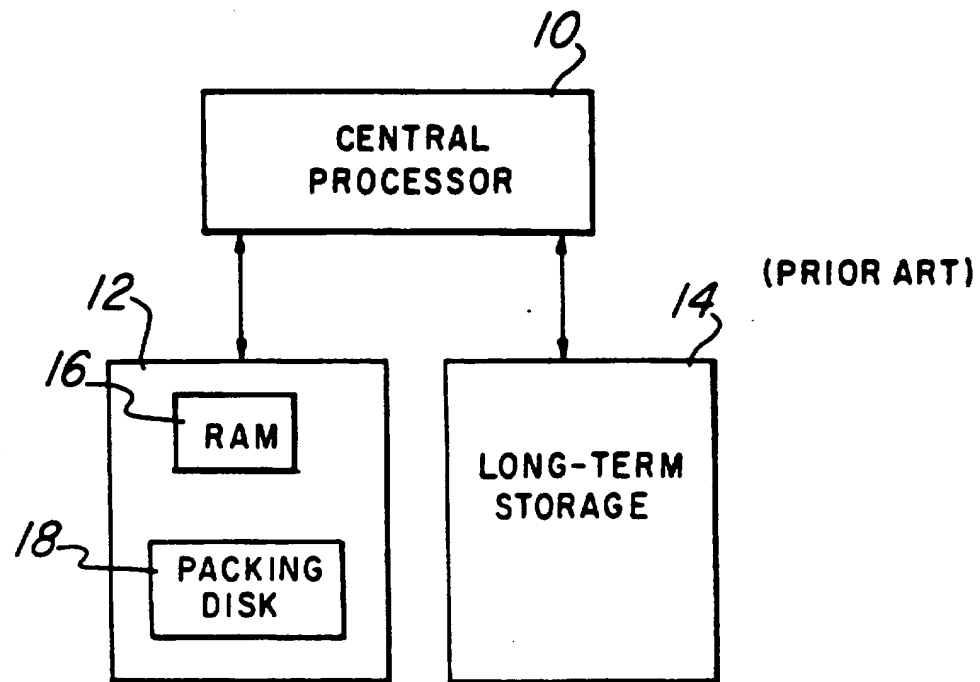
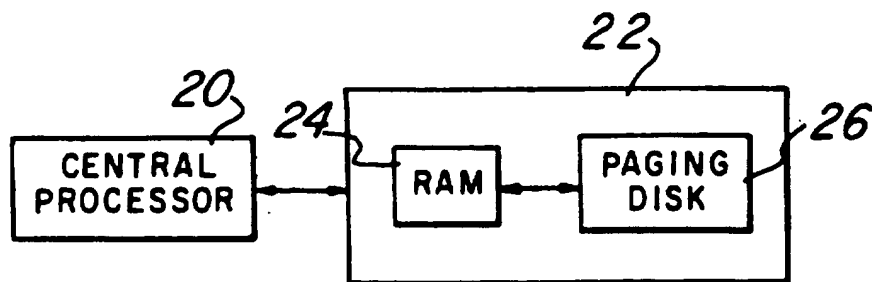
Attorney, Agent, or Firm—L. Joy Griebenow; James T. Comfort; Melvin Sharp

## [57] ABSTRACT

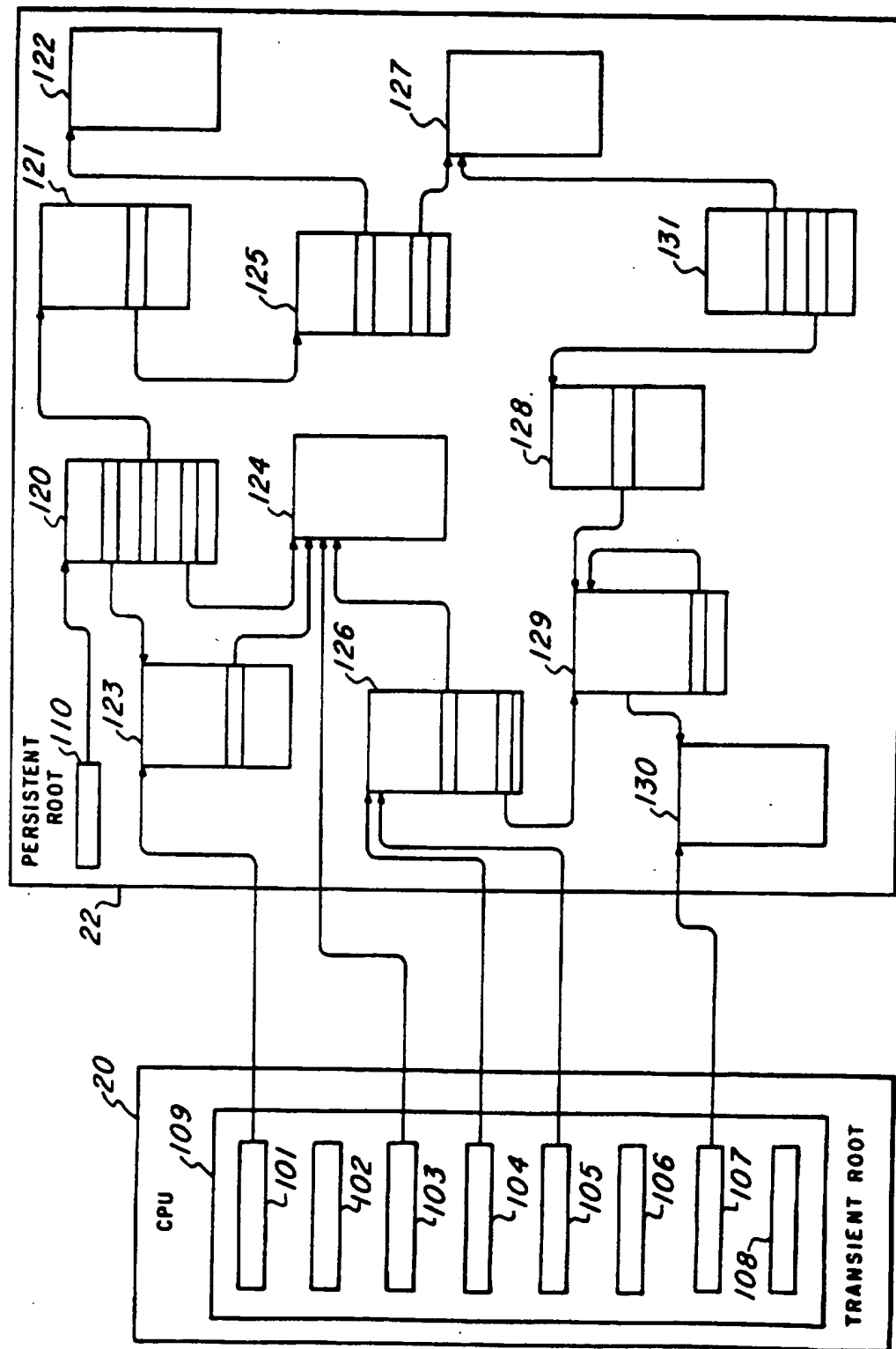
A recoverable virtual memory for a computer system takes periodic checkpoints which capture the state of the virtual memory. If a system failure occurs, the system can be rolled back to the checkpointed state and restarted. A mechanism for tracking which virtual pages are contained in the checkpointed state discards pages which have been modified since the checkpointed state was saved. Only versions of pages which are saved in the checkpointed state are used in the restore process.

10 Claims, 6 Drawing Sheets

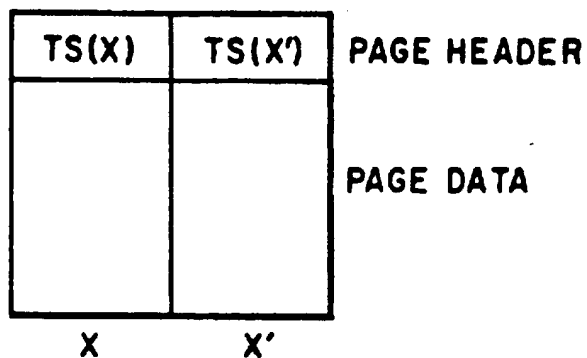
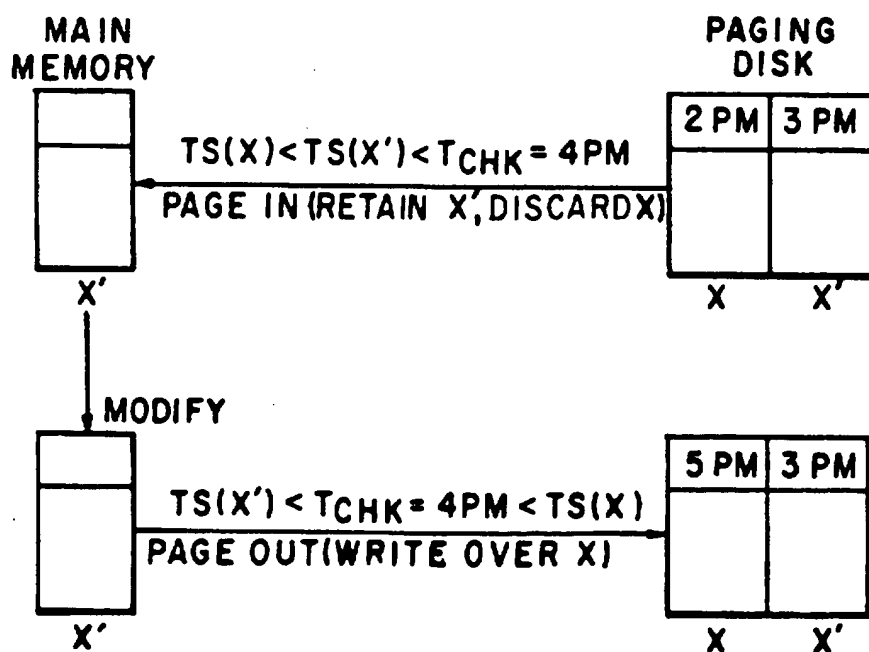


*Fig. 1**Fig. 2*





**Fig. 3**

*Fig. 4**Fig. 5*

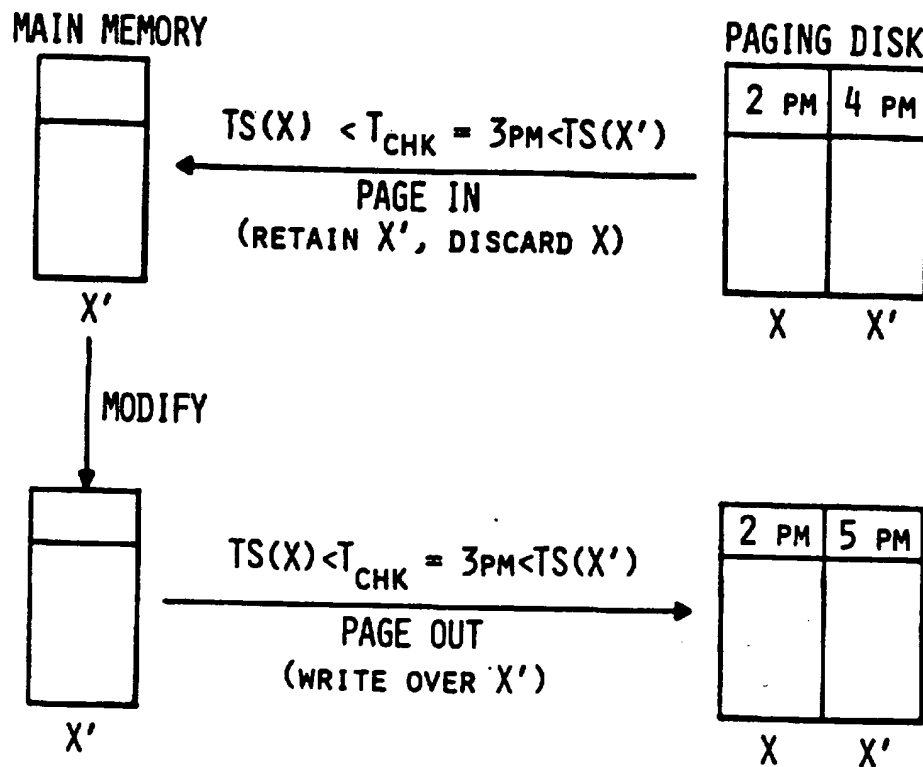


Fig. 6

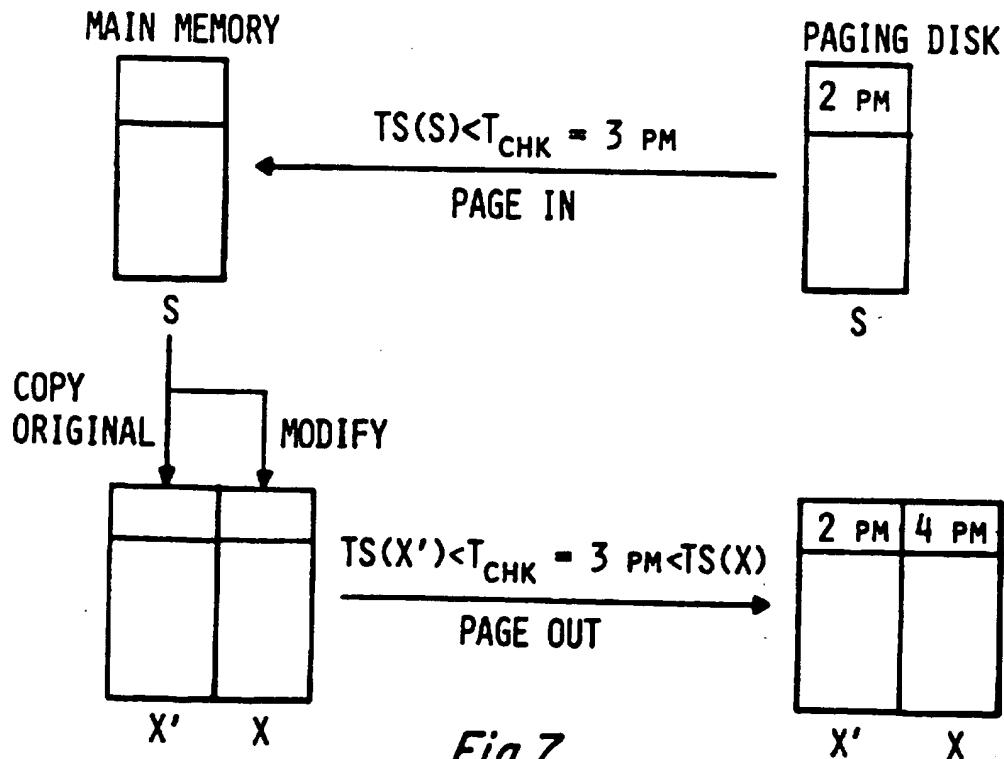
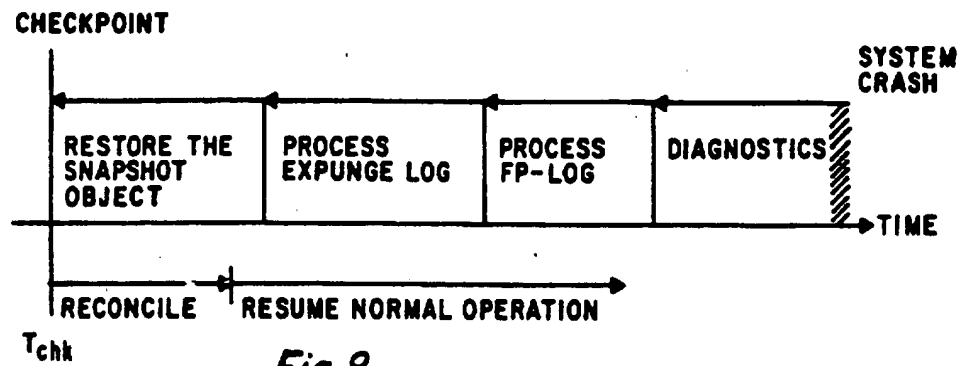
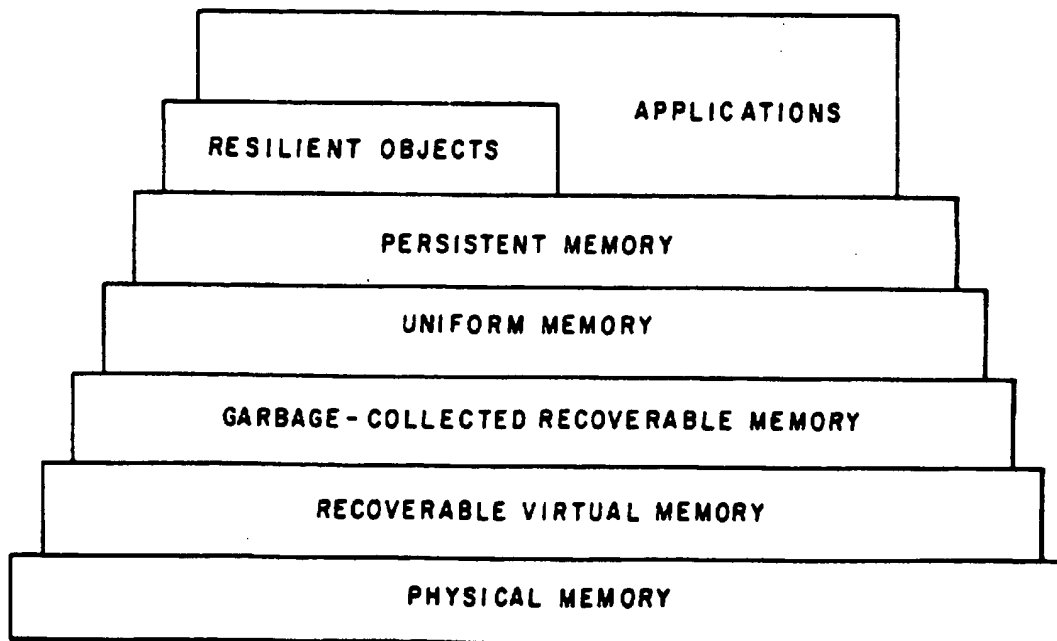


Fig. 7

*Fig. 8**Fig. 9*

VIRTUAL ADDRESS <sub>0</sub>	DISK ADDRESS <sub>0</sub>	OTHER <sub>0</sub>	STATE <sub>0</sub>
VIRTUAL ADDRESS <sub>1</sub>	DISK ADDRESS <sub>1</sub>	OTHER <sub>1</sub>	STATE <sub>1</sub>
⋮			
VIRTUAL ADDRESS <sub>N</sub>	DISK ADDRESS <sub>N</sub>	OTHER <sub>N</sub>	STATE <sub>N</sub>

*Fig.10*

## RECOVERABLE VIRTUAL MEMORY HAVING PERSISTANT OBJECTS

The present application is a continuation-in-part of U.S. Ser. No. 774,828, titled RECOVERABLE VIRTUAL MEMORY filed Sept. 11, 1985, by Satish Thatte.

### BACKGROUND AND SUMMARY OF THE INVENTION

This invention relates to digital computer systems, and specifically to a virtual memory having recovery capabilities.

In the future, as users of state-of-the-art symbolic computing machines develop large-scale, knowledge-based applications, they are expected to encounter major problems arising out of storage management problems in supporting large and complex knowledge/data bases. The word storage is used herein in a broad sense to encompass virtual memory, file systems and databases. The problems can be primarily attributed to the dichotomy by which today's computers, including state-of-the-art symbolic computers such as the Texas Instruments EXPLORER and the Symbolics 3670, manage storage along two entirely different organizations. These organizations can be referred to as the computational storage and the long-term storage.

In symbolic/artificial intelligence (AI) processing, a representation of knowledge is a combination of data structures and interpretive procedures that, if used in the right way in a program, will lead to "knowledgeable" behavior. The goals of AI systems can be described in terms of cognitive tasks like recognizing objects, answering questions, and manipulating robotic devices. The most important consideration in formulating a knowledge representation scheme is the eventual use of the knowledge. The actual use of the knowledge in symbolic/AI programs involves three stages: (1) acquiring more knowledge, (2) retrieving facts from the knowledge base relevant to the problem at hand, and (3) reasoning about these facts in search of solutions. A number of different knowledge representation schemes, such as state-space representation, logic, procedural representation, semantic nets, production systems, and frames, have been developed by the knowledge representation community. The choice of the knowledge representation scheme very much depends on the application requirements.

No matter which knowledge representation scheme is used, at some sufficiently low level of representation the knowledge is represented by memory objects interconnected by pointers. These objects exhibit a structure which is defined by the interconnection graph of pointers connecting the objects. The structure of objects created and manipulated by symbolic AI applications is usually very rich and complex. Moreover, both the information in objects, as well as the structure of objects, can undergo rapid changes.

In symbolic computing, objects representing a knowledge base are created and manipulated in the computational storage. As its name implies, the computational storage contains objects to be manipulated by the processor of a computer system. These objects can be numbers, strings, vectors, arrays, records, linked lists, instructions, procedures, etc. These objects, both small and large, are usually identified by names. The names of objects serve as convenient handles or point-

ers that can be passed as procedure parameters, returned as procedure results, and stored in other objects as components. The names of objects are typically implemented as their virtual addresses. Programmers create and manipulate objects by using programming languages, such as LISP and PROLOG.

Typically, the computational storage is implemented as virtual memory, which consists of a hierarchy of memories: a fast, small semiconductor main memory, backed up by a slow, large disk to support paging. Objects in the computational storage are accessed very rapidly as the processor can directly access them by specifying their addresses (real or virtual), often at a speed that matches the basic processor cycle time. The information stored in these objects is also processed and manipulated very efficiently as it is stored in a format defined by the processor architecture, and can therefore be directly interpreted by the processor hardware or microcode.

Often, the information stored in the computational storage has a very rich structure; i.e., objects in the computational storage are interconnected by a rich and complex structure of pointers to match the requirements of applications at hand. The structure of these objects is often dynamic. However, objects in the computational storage do not exist beyond the life times of programs that create them. When a program terminates or a system shutdown, or crash occurs, these objects cease to exist. Therefore, they are called short-lived or transient objects. To make these objects survive beyond the life times of programs that created them, i.e., to make them long-lived or persistent, they must be moved to the other storage organization, i.e., the long-term storage.

As its name implies, the long-term storage is used to keep information for long periods of time. It is typically implemented on a disk-resident file system. The disk file system is logically different from the paging disk of the computational storage, even though the physical disk media may be shared by both. Examples of information stored in the long-term storage are files, directories, libraries, and databases. The long-term storage retains information in a reliable fashion for long periods of time. In order to store information beyond the life time of a program that periods of time. In order to store information beyond the life time of a program that creates it in the computational storage, the information needs to be first mapped into a representation expected by the long-term storage and then transferred to it for long-term retention using a file input/output (I/O) operation or a database operation.

The types of objects supported by the long-term storage are very restrictive (essentially files, directories, relations, etc.), and may match with the data structure requirements of many applications. The representation of information in the long-term storage is quite "flat." For example, a file may consist of a sequential stream of bits or bytes, such as ASCII characters. Files or relations usually can neither hold procedural objects nor pointers to other objects in the long-term storage. Information in these objects can neither be directly addressed nor directly processed by the processor, because its representation is not compatible with the processor architecture. The information can be processed only after it is mapped into a representation expected by the computational storage and then transferred to it for processing. The translation overhead in mapping these

objects to/from a collection of files is quite substantial, too.

In addition to the time overhead for translation and mapping of objects between the computational and long-term storages, there is additional space overhead, as the information is essentially duplicated in virtual memory and the file system. There is an apparent paradox in that the computational storage, usually implemented as a virtual memory, hides the existence of the paging disk store; on the other hand, the long-term storage makes the existence of the disk explicit to the programmer. Thus, the programmer is faced with a nonuniform storage model, where differences in addressing, function, and retention characteristics between the computational and long-term storages are visible above the processor architecture level.

Programming languages, such as FORTRAN, Pascal, LISP, and Prolog, strongly reflect the dichotomy in storage organization. The specification of these languages almost invariably assumes long-term storage objects (files) to have entirely different characteristics from computational objects. As a result, these programming languages cannot directly process information in the long-term storage the way they can process information in the computational storage. This dichotomy propagates throughout the whole system and cannot be hidden from the user. It shows up in differences between names used for programming language objects and names used for files and databases.

The dichotomy also shows up in a different set of languages that has evolved to process information in the long-term storage. These languages include various so-called command languages, such as the UNIX shell language and the IBM TSO Command Language, that are responsible, among other things, for performing operations on files. The other class of languages which operate on persistent objects are various database languages, such as Square, Sequel, and Quel. These languages can define database objects, and perform queries and updates on them. Typically, such languages are often interpreted, and are restrictive and arcane in nature compared to the more familiar programming languages, which also enjoy the efficiency of compiled execution over interpreted execution.

As a consequence, the programmer must be aware of the nonuniform storage model, and must explicitly move information among storage media, based on the addressing mechanisms, functions and retention characteristics desired. Another consequence is that the nonuniform storage model is an obstacle to programming generality and modularity as it increases potential types of interfaces among programs. The hodgepodge of mode-dependent programming languages, such as command languages, programming languages, debugging languages, and editing languages, makes fast and efficient interaction with the system difficult.

The mapping between transient and persistent objects is usually done in part by the file system or the data base management system (DBMS) and in part by explicit user translation code which has to be written and included in each program. This task imposes both space and time penalties, and degrades system performance. Frequently the programmer is distracted from his task by the difficulties of understanding the mapping and managing the additional burden of coping with two disparate worlds: the programming language world and the DBMS world.

In large data-intensive programs there is usually a considerable amount of code, which has been estimated to be as high as 30% of the total, concerned with transferring data between files or a database, and the computational storage. Much space and time is wasted by code to perform translations between the transient and persistent object worlds, which has adverse performance impact. This is unsatisfactory because the effort and time required to develop and execute the translation code can be considerable, and also because the quality and reliability of the application programs may be impaired by the mapping. The storage dichotomy also gives rise to much duplication of effort in the operating system design and DBMS design.

These problems, created by the storage dichotomy, are considerably further complicated for symbolic/AI computing. Processes on current symbolic machines share a single address space: i.e., there is no per-process address space. Moreover, the address space is not segmented, but is a single, linear address space. Such a model of the computational storage allows easy, efficient and flexible sharing of objects among multiple processes. Any object can point to any other object by simply holding a pointer to that object (usually implemented as a virtual address of the object being pointed to). Arbitrarily complex structures of objects interconnected by pointers can be created and manipulated. Such powerful structuring of objects is very important for the development of the highly integrated and powerful software development environments available on these symbolic computers.

Unfortunately, current symbolic computers make a distinction between the computational and long-term storages, similar to today's conventional computers. In symbolic computers, making a single object persistent by moving it to a file system is not very meaningful; all objects that can be reached from an object by following all out-going pointers also need to be made persistent as a single entity, and all in-coming pointers pointing to the entity must be "properly taken care of." Such an entity, however, can be very large and moving it to a file system would be a complicated and expensive operation. Conversely, the reverse move from a file system to the computational storage would be equally as complicated and expensive.

Many current advanced programming techniques, especially as practiced in the symbolic/AI community, do not distinguish between procedures and data; procedures are just data, which are themselves active. As the body of information being dealt with grows and becomes more active, it becomes critical that the program environment, consisting of complex objects interconnected with rich pointer structures, survives for long periods of time. Mapping and moving of such rich environments into today's file system or database for long-term retention would involve substantial translation overhead, both in space and time.

Thus, there is a substantial difference between the representations of objects in the computational and long-term storages for symbolic/AI applications. The richer the structure of computational objects, the greater the difference and the bigger the effort needed to perform translation between these two representations. Emerging symbolic and AI applications will employ increasingly sophisticated and complex structures on a large number of objects on which retrievals, queries, inferences, reasoning, deductions, and computations will be performed. As can be anticipated, the

overhead to map long-term objects into computational objects and vice-versa for large knowledge-intensive applications could be substantial.

The current approach taken by many researchers to facilitate knowledge-based applications is based on connecting a symbolic computer to a database machine. This approach is not based on persistent memory, as it neither addresses the storage dichotomy issues nor deals with the lifetime or interchangeability of procedure and data issues. There will be a mismatch between the data model requirements of symbolic/AI applications and the rigid data models supported by database machines. Therefore, such approach appears to be inadequate for expert database systems. These reservations are shared by other researchers in the field.

The persistent memory approach is based on a fundamentally different foundation. The literature on persistent memory dates back to 1962, when Kilburn proposed single-level storage, in which all programs and data are named in a single context. (T. Kilburn, "One Level Storage System", IRE Trans. Electronic Comput., vol. EC-11, no. 2, Apr. 1962) Saltzer proposed a direct-access storage architecture, where there is only a single context to bind and interpret all objects. (J. H. Saltzer, "Naming and Binding of Objects". In R. Bayer et al, editors, Operating Systems, An Advanced Course, p. 99, Springer-Verlag, New York, N.Y., 1978.

Traiger proposed mapping databases into virtual address space. (I. L. Traiger, "Virtual Memory Management for Database Systems", ACM Operating Systems Review, pp. 26-48. Oct. 1982.) It seems that the simple data modeling requirements of the FORTRAN and COBOL worlds discouraged productization of these proposals because they are much more difficult to implement than the conventional virtual memory and database systems.

The MIT MULTICS system and the IBM System/38 have attempted to reduce the storage dichotomy. However, both have major shortcomings for symbolic computing: unlike LISP machines, each process has its own address space. All persistent information is in files. A file mapped into the address space of a process cannot hold a machine pointer to a file mapped in the address space of a different process. Thus, sharing of information among different processes is more difficult than with LISP machines. Furthermore, there is no automatic garbage collection, which is essential for supporting symbolic languages.

Recently, many researchers have proposed implementing persistent objects on top of a file system provided by the host operating system. Though persistent and transient objects still reside in two separate storage organizations, persistent objects can be of any general type, such as number, vector, array, record, or list, and can be manipulated with a common programming language such as ALGOL or LISP. However, there is a large overhead to access persistent objects because their pointers must be dereferenced by software, taking several machine cycles.

Systems having two-level memory storage can easily recover from a power failure, hardware failure, software error, or the like, which can be considered as a group as "system crashes". After a system crash, any hardware problems are repaired and the software is reloaded from long-term storage. All data and procedures which were in the virtual memory at the time of the crash are discarded, and the system is restarted, and

those items that have been stored in files or a DBMS are considered to be valid.

A system which implements a large uniform memory is especially vulnerable to system crashes. Because persistent objects are stored in the virtual memory, they can be corrupted by the crash. The most recent version of a particular persistent object may or may not be stored on the paging disk. The current value of large objects may be partially on disk, and partially in RAM. Thus, the values stored on disk cannot be relied on, and cannot merely be used to reload and restart the system after a crash.

Thus, if it is desired to restore a virtual memory after a crash, prior art file and DBMS systems cannot be used. It is necessary to devise some mechanism for preserving the state of the virtual memory.

It is an object of the present invention to provide a virtual memory which can recover from hardware failures and software errors. It is a further object to provide a virtual memory which can be restored to an earlier, valid state to minimize loss of work. It is another object to provide a means for taking regular checkpoints of the virtual memory to preserve valid states which can be restored. It is another object to provide an improved recoverable paging scheme for virtual memories.

In order to provide for system recovery in case of a power failure, hardware failure or software error, checkpoints are periodically taken of the state of the system. These checkpoints are marked and stored on disk. Changes made between a checkpoint and the next checkpoint are also stored and marked, but are discarded in the event of a system failure. When there is a system failure, the system is rolled back to the checkpoint state, and processing resumes in a normal manner. Virtual memory pages which are updated after the most recent checkpoints are stored on disk as sibling pages. An efficient state indicator mechanism is provided for determining which sibling page is to be read from or written to disk when the corresponding virtual page is referenced. This state indicator mechanism indicates which pages on disk are included in the checkpoint state, and which contain information modified since the checkpoint. In normal operation, the most recent version is used, but when a system failure occurs, only the most recent version stored before the checkpoint is used.

The novel features which characterize the present invention are defined by the appended claims. The foregoing and other objects and advantages of the invention will hereinafter appear, and for purposes of illustration, but not limitation, two preferred embodiments are shown in the drawings.

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a prior art computer memory system architecture;

FIG. 2 is a block diagram of a computer memory architecture for implementing a persistent memory;

FIG. 3 is a block diagram illustrating a uniform, persistent system;

FIG. 4 is a representation of sibling virtual pages as stored on a paging disk according to a first preferred embodiment;

FIG. 5 illustrates how some sibling pages are updated according to a first preferred embodiment;

FIG. 6 illustrates how other sibling pages are updated according to a first preferred embodiment;



FIG. 7 illustrates how singleton pages are updated according to a first preferred embodiment;

FIG. 8 illustrates the process of recovering from a system crash;

FIG. 9 illustrates a hierarchy of memory abstractions; and

FIG. 10 is a page table according to a second preferred embodiment.

## DESCRIPTION OF THE PREFERRED EMBODIMENT

The present recovery system for a virtual memory will be described in the context of its implementation with a single-level persistent memory, which will be described first. It will be apparent to those skilled in the art that the present recoverable virtual memory can be used with other virtual memories, such as those used with present virtual memory systems.

### PERSISTENT MEMORY SYSTEM

FIG. 1 shows a prior art computer architecture having separate computational and long-term storages. A central processor 10 has access to computational storage 12 and long-term storage 14. Long-term storage 14 is for retaining files, databases, etc., and is usually implemented as one or more disks backed up by magnetic tape. Computational storage 12 is a virtual memory, usually implemented as a fast semiconductor RAM memory 16 and a paging disk 18. The computational storage 12 appears to the central processor 10 as a very large RAM. Virtual memory addresses which are not actually in the semiconductor memory 16 are located on the paging disk 18 and loaded into the semiconductor memory 16 when they are referenced by the central processor 10.

FIG. 2 shows a computer system having an architecture according to the present invention. A central processor 20 (CPU) has access only to a single, uniform memory 22. The memory 22 preferably consists of a very large virtual memory, having semiconductor RAM 24 backed up by a paging disk 26. This virtual can be very similar to the virtual memory 12 of FIG. 1. The CPU 20 may be an existing system, such as an EXPLORER symbolic processor from Texas Instruments. The virtual memory 22 appears to the CPU 20 as a uniform, or single-level, memory store with a linear address space.

The uniform memory abstraction defines the storage system architecture necessary to implement a persistent memory according to the present invention. The persistent memory system is based on the uniform memory abstraction, in which a processor views memory as a set of variable-sized blocks, or objects, of memory interconnected by pointers. The memory system has a very large address space to support large knowledge-based applications. The persistent memory system is expected to store persistent objects, including "files," which could be very large in number and size. Therefore, the size of the underlying address space should be sufficiently large to support a practical system. However, the concept of persistent memory does not depend on the actual size of the address space.

As previously explained, all processes within a symbolic computer share the same single, linear address space. This allows a rich, complex structure of objects interrelated by pointers, to be created and manipulated. The structure of memory objects interconnected by pointers forms a graph. Pointers interconnecting mem-

ory objects are implemented as virtual addresses of the target objects.

As shown in FIG. 3, there is a distinguished object in the uniform memory abstraction, called the persistent root 110, which defines persistent objects.

The persistent root 110 is a distinguished object located at a fixed virtual address and disk location. All objects that are in the transitive closure of the persistent root, i.e., reachable from the persistent root by following pointers, are persistent. The persistent root survives system shutdowns or crashes. Typically, the persistent root may contain a pointer to a table that points to other tables or structures of persistent objects and so on. Thus, the persistent root anchors all persistent objects.

The persistence attribute of an object depends solely on whether that object can be prevented from being garbage collected even after the program that created it has terminated: this can be easily arranged by making that object a member of the set of objects in the transitive closure of the persistent root. Persistence based solely on the persistent root rather than the properties of the storage medium allows a complete separation of the persistence attribute of an object from its type or relationship with other objects. Numbers, characters, lists, procedures, environments, etc., can be persistent objects while they exist in virtual memory.

Therefore, an invocation of a procedure as a persistent object is as easy and efficient as its invocation as a transient object. In fact, from the machine point of view, transient and persistent objects are indistinguishable. From the user point of view, there is no need to treat transient and persistent objects differently; all the user needs to know is that to make an object persistent, it has to be in the transitive closure of the persistent root.

The processor contains a number of "registers." (101-108 are shown) The processor can access a memory object, i.e., read and write its individual words, if any of its registers holds a pointer to the object. The word register in this context is used in a generic sense; it may be a hardware register or a scratch-pad memory in the processor. These registers define the transient root 109 of the memory system. They do not survive a system shutdown or crash. All objects that are in the transitive closure of the transient root, but not in the transitive closure of the persistent root, are called transient. All the remaining objects are garbage and are reclaimed by a garbage collector.

FIG. 3 shows an example snapshot of the memory system and categorizes objects within it. The arrows between objects and from CPU registers to objects represent pointers. Pointers always refer to the beginning of the object pointed to. Thus, the four pointers pointing into object 124, for example, all have the same value and point to the beginning of block 124. By determining the transient closure of the persistent root 110, and the transient root 109, it is seen that objects 126, 129 and 130 are transient; objects 120, 121, 122, 123, 124, 125, and 127 are persistent; and objects 128 and 131 are garbage.

Each memory object consists of one or more memory words, or cells, which are stored in consecutive virtual addresses. The processor 20 can access a memory object, i.e., read and write its individual words, if any of its registers holds a pointer to the object. For example, one method of accessing individual cells is as follows. If register 101 contains a pointer to a memory object 123, then the processor 20 can read the third word of the

memory object 123 by executing a READ(1, 3) instruction, where "1" specifies the processor register 101, and "3" specifies the third word of the memory object (123), pointed to by register 101. The contents of register 101 are added to "3" to develop the virtual address of the word to be read. Similarly, the processor 20 can write data in the fourth word of the memory object 123 by executing a WRITE(1, 4) data instruction. The processor 20 can access memory objects only via logical addresses; a logical address consists of a pair (i, j), where "i" is the identification number of a processor register, and "j" indicates the j-th word of an object being pointed at by processor register "i."

The notion of memory objects in the uniform memory abstraction corresponds to objects used in high-level programming languages, such as numbers, booleans, characters, strings, LISP CONS cells, arrays, records, procedures, or environments. These language-level objects can be implemented using one or more memory objects interconnected by pointers. Application-level objects are constructed by combining language-level objects.

The persistence property of objects is based solely on whether or not an object is within the transitive closure of the persistent root 110. The persistence attribute of an object is a fundamental notion. It should depend only on whether the object can survive beyond the life time of a program that creates it. It should neither depend on the type of the object nor on the properties of the storage medium on which the object resides. Since there will usually be several sets of unrelated groups of objects which are persistent, the persistent root 110 will usually first point to an object which contains nothing more than pointers to persistent objects. Any object, transient or persistent, can point to any other object to facilitate the unrestricted sharing desired in many symbolic/AI computations.

In contrast to this mechanism of achieving persistence of objects based solely on the persistent root, in today's machines, both conventional and symbolic, an object becomes persistent only when it is stored in the long-term storage, i.e., disk store. Even in MULTICS or IBM-System/38, only certain types of objects, i.e., files, can become persistent, while other types of objects, such as procedures, cannot.

With the persistent root 110, the persistence attribute of an object solely depends on whether that object can be prevented from being garbage collected even when the program that created it has terminated; this can be easily arranged by making that object a member of the set of objects in the transitive closure of the persistent root 110.

The transience/persistence attribute of objects is not necessarily a permanent attribute. An object may be created as a transient object, then it can become a persistent object solely on the basis of being in the transitive closure of the persistent root 110, and then can revert back to the transient state by getting out of the transitive closure of the persistent root 110, and so on.

Each pointer, implemented as a virtual address, is tagged as a pointer within memory. This tagging mechanism is used to ensure that the processor cannot specify, fabricate, or forge a pointer. The processor is allowed to access memory only by reference to logical memory blocks. There may be additional tagging information associated with each object to indicate its type, such as integer, floating point number, string, array, list, or closure. This tagging information is used to ensure

that attempts to perform operations that are undefined or illegal on a particular object type cause traps to appropriate exception handling routines; for example, an attempt to add an integer to a string object would cause an exception. Each memory reference can be checked for bounds, i.e., "j" in a logical address (i, j) should not exceed the size of the object pointed to by processor register "i."

The nature of the memory system requires that it be garbage collected and be free from the so-called dangling reference problem. Garbage collection is essential to be able to make computational progress in a finite amount of memory space. Without the reclamation and reuse of memory space occupied by an object proven to be garbage (i.e., no outstanding pointers to the object from non-garbage objects), the system would eventually come to a grinding halt as it ran out of memory. Garbage collection is preferably done automatically in real time, and preferably as a process executing concurrently with user processes. This is not necessary to the invention, however, and garbage collection can occur during periods of machine non-use, such as overnight.

The dangling reference problem arises if the memory space for an explicitly deleted object is reclaimed without proving that there are no outstanding pointers to that object. If the space occupied by a deleted object is reclaimed prior to such a proof, then the outstanding pointers to the object may point to empty space, i.e., unallocated memory, or to some undesired object if the reclaimed space has been later allocated to the new object. In either case, the memory system integrity would be violated.

The proof that there are no outstanding pointers to a deleted object is embedded within the garbage collector. A deleted object is specially marked as tombstoned when it is deleted. All of the outstanding pointers to a tombstoned block are detected by the garbage collector and then destroyed. When no pointers to a tombstoned block are left, its space is automatically reclaimed.

The notion of the integrity of the memory system refers to the integrity of the structure of the graph of memory objects interconnected by pointers. Under this notion, the integrity of the memory system is preserved because it is impossible to forge a pointer, no information outside the bounds of an object can be accessed, undefined or illegal operations for any object type cause appropriate exception handling actions, and there is no dangling reference problem. Therefore, the graph of memory objects undergoes transition only from one consistent state to another, ensuring its structural integrity.

The uniform memory abstraction achieves its goal of being able to manipulate both transient and persistent objects with equal ease, flexibility and efficiency because words of a memory object pointed to by a processor register can be accessed with equal ease and speed, independent of whether the object is transient or persistent. Therefore, an invocation of a procedure as a persistent object is as easy and efficient as its invocation as a transient object. In fact, as far as the machine is concerned, there is no distinction in processing transient and persistent objects. The uniform memory abstraction allows the same set of flexible and powerful operations on transient and persistent objects from a single programming language. New procedures can be invoked as persistent objects or transient objects. None of the prior art systems can do this; in those systems, access to per-

sistent objects is considerably slower or more complicated than access to transient objects.

The relationship between the parts of the overall memory system, including the uniform and persistent memory, just described, and the recoverable virtual memory, described below, is illustrated in FIG. 9. The various levels of viewing the system are shown as a hierarchy of memory abstractions.

At the lowest level is the actual physical memory: the main semiconductor RAM memory and the paging disk. The next level of abstraction is the recoverable virtual memory as described below. The next level shows garbage collection of the virtual memory, which can be done by one of several known methods. The abstraction of uniform memory is one of several that can be implemented on top of a recoverable virtual memory. The persistent memory described above is implemented as the next layer of abstraction, and the concept of resilient objects, as described below is implemented on the persistent memory. Finally, user applications, such as expert database systems, are implemented at the top of the hierarchy.

#### RECOVERABLE VIRTUAL MEMORY

The recoverable virtual memory will now be described. In order to preserve the contents of the persistent memory system in the event of a system crash, a recovery capability at the level of the virtual memory itself is required. In today's machines, both conventional and symbolic, recovery capability at the level of virtual memory is not required. If there is a system crash, information in the computational storage is assumed to be lost; but information in the long-term storage survives because it resides on a non-volatile disk store, which cannot be directly accessed by the processor. Applications or users are expected to transfer information from the computational storage to the long-term storage for survival from crashes. After a system crash, the system is restarted using the information found in the long-term storage. Therefore, recovery is not a major concern in today's machines. Such a luxury has to be sacrificed in the persistent memory system, as there is no distinction between the computational and long-term storage.

Failures that are expected to be dealt with by the recovery system can be classified into two broad categories: system crashes and media failures. System crashes may be caused by power failures, or software errors or hardware failures. Software errors may arise from bugs in the application or systems software, incorrect data or commands, etc. Hardware failures account for intermittent and permanent failures in the processor, main memory, or I/O devices and their controllers. Media failures, which account for disk storage failures, have been deliberately excluded from hardware failures for reasons explained below.

A system crash can be signaled by hardware self-checking circuits, a power failure detector interrupt, or by the inability of software error handling routines to recover from a software error. Hardware self-checking circuits typically include parity error detectors on memories and buses, fire code detectors for disks, self-checking checkers, watch-dog processors, etc. It is assumed that the latency of a system crash detection, i.e., the time period between the system state corruption because of a system crash and the actual detection of the crash, is quite small so that the probability of taking a

checkpoint, as described below, in this small time period is negligibly small.

When a system crash occurs, the present state of the entire uniform memory is assumed to be corrupted and no further computational progress is possible. When a system crash is detected the diagnostics routines are run, which attempt to identify if the failure occurred due to a permanent hardware failure. If a permanent hardware failure is detected, then the faulty hardware, such as a processor board or a memory module, is replaced or repaired before any further recovery steps are taken. If no permanent hardware problem is detected, the recovery scheme rolls back the state of the entire memory system to the last checkpoint state stored on disk, as explained below. From the checkpoint state, normal computation can be resumed.

It is expected that system crashes arising from software failures will be much more common than those arising from hardware failures. The average rate of system crashes in a single-user machine is expected to be on the order of a few crashes per month, and an average recovery time of several minutes is acceptable (assuming no permanent hardware failures).

The media failures that affect disk storage can arise from disk head crashes, deterioration of the magnetic media itself, or bugs in the disk driver software. It is expected that media failures are much less frequent than system crashes. The average media failure rate may be a few failures per year, and the average recovery time after a media failure is expected to be a few hours.

The reason that media failures are treated differently from system crashes is that a media failure also corrupts the last checkpoint state on disk, and the roll-back technique used for recovery after a system crash would not work. To deal with media failures, the last checkpoint state on disk needs to be archived on another media, such as a streaming tape. After a media failure, a failed disk needs to be replaced with a new one, which is then initialized from the last archived checkpoint state.

A preferred recovery scheme for recovering from system crashes, as defined above, will now be described. The recovery scheme is based on a novel mechanism to take periodic checkpoints of the entire machine state in an efficient fashion, possibly initiated by the user or an application program. The checkpointing scheme keeps the entire machine state valid, within the last few minutes, on disk.

The preferred embodiment utilizes a paged virtual memory. Memory is paged to and from the paging disk in blocks of predetermined size, for example two thousand words, regardless of the size of memory objects 120-131. Thus, several small objects may reside on a single page, and large objects may extend over multiple pages.

When a new virtual page is materialized at the request of the memory allocator, it is placed in a properly initialized main memory 24, and is simultaneously materialized on the paging disk 26. As explained below, a virtual page is materialized on disk in either the sibling form or the singleton form. In the preferred sibling form, two adjacent disk blocks are allocated to a virtual page; in the singleton form, a single disk block is allocated. If the page is unlikely to be modified in the future (for example, a page containing instructions), it is materialized in the singleton form to save disk space. If it is expected to contain data, it is materialized in the sibling form.

As shown in FIG. 4, the two adjacent disk blocks allocated for a virtual page in the sibling form are called siblings. The sibling pages are referred to as  $x$  and  $x'$ . A singleton page is denoted as  $s$ .

As shown in FIG. 4, the page header is contiguous to the page data on disk. As a page is fetched from the paging disk 26 into main memory 24, the header information is scattered to some other predetermined page in main memory 24. When the page is written to disk, its header information is gathered together with the data. Scattering refers to the act of separating the page header from the data, and gather refers to the reunion of the header and data. This is done because the data portion of the page corresponds to a page in main memory, and the header is additional information. To perform this operation efficiently, the disk controller should have the scatter-gather capability. The scatter-gather capability is necessary only for performance and not for the correct functionality of the recoverable virtual memory.

When a page is written to the paging disk 26, it is stamped with the time of the disk write operation. In FIG. 4,  $TS(x)$  indicates the timestamp of  $x$  and  $TS(x')$  indicates the timestamp of  $x'$ .  $TS(s)$  denotes the timestamp of  $s$ . A convenient place to record the timestamp is the page header. Timestamps written in headers will survive system crashes. Timestamps are derived from a timer that runs reliably even in the presence of system crashes. The granularity of timestamps need only be moderately smaller than the time for a disk write operation because pages cannot be written faster than the disk write speed. With a 10 milliseconds granularity, a 64 bit timer can generate unique timestamps for 5.8 billion years. Therefore, a 64-bit timestamp field in page headers is more than adequate.

When a page is materialized in sibling form, its siblings are materialized on disk with timestamps of  $-1$  and  $-2$ , indicating that both are yet to be written. Alternate schemes are of course possible. When a page is materialized in singleton form, it is initialized on disk with a time stamp of  $-1$ . All disk blocks that have been modified since their initial materialization on disk will have unique timestamps.

The time of the last checkpoint operation is denoted as  $T_{chk}$ . It is stored in a reliable fashion at a known disk location. Updating of  $T_{chk}$  must be assured to be reliable; any errors made in updating  $T_{chk}$  will probably destroy the checkpoint scheme. One method for ensuring the integrity of the value  $T_{chk}$  is to keep two identical copies. When  $T_{chk}$  is to be updated, the first copy is changed, then checked to ensure a proper update. A system crash during this operation would not effect the second copy, so that the original value of  $T_{chk}$  is preserved. After the first copy is updated, the second copy is updated and checked. A system crash during this operation will not effect the system integrity, as the proper new value of  $T_{chk}$  is in the first copy. Thus, one is assured of a proper value for  $T_{chk}$  at all times.

A checkpoint state is "pristine" in the sense that it cannot be accessed during normal operation, and remains unperturbed unless there is a media crash. A singleton page with a timestamp smaller, or earlier, than  $T_{chk}$  belongs to the checkpoint state; on the other hand, a singleton page with a timestamp larger, or later, than  $T_{chk}$ , is outside the checkpoint state. For sibling pages, if  $TS(x) < TS(x') < T_{chk}$ , i.e., if both siblings have timestamps smaller than  $T_{chk}$ , then the sibling with the larger timestamp, here  $x'$ , belongs to the checkpoint state, and

the other sibling, here  $x$ , contains outdated information. If one sibling has a timestamp smaller than  $T_{chk}$  and the other sibling has a larger timestamp, the sibling with the smaller timestamp belongs to the checkpoint state and the sibling with the larger timestamp is outside the checkpoint state. Because of the way that timestamps are initialized and updates, both siblings cannot have a timestamp larger than  $T_{chk}$ . The timestamp is the only mechanism to distinguish between pages that belong to the last checkpoint state and pages that do not.

On a page fault, if the missing page is in the sibling form, both sibling pages are read in a single disk access to reduce access time. This is possible because they are located adjacent to each other on disk. It is expected that the time to fetch two adjacent disk pages is only marginally larger than the time to fetch a single page. For the TI EXPLORER LISP machine, the average additional time to fetch the second disk page is expected to be less than approximately 2% of the total.

If  $TS(x) < TS(x') < T_{chk}$ , i.e., if both siblings have timestamps smaller than  $T_{chk}$ , then the sibling with the larger timestamp, i.e.,  $x'$ , is kept in main memory, and the other sibling, i.e.,  $x$ , is discarded. The situation is analogous for  $TS(x') < TS(x) < T_{chk}$ . The sibling kept belongs to the checkpoint state, while the discarded sibling contains useless information. When this page is later written to disk, it is written over the disk space of the discarded sibling  $x$ : it cannot be written over the disk space of  $x'$ , else the checkpoint state would be destroyed. After writing again to disk, the timestamp relationship becomes  $TS(x') < T_{chk} < TS(x)$ . This operation is illustrated in FIG. 5, which assumes the most recent checkpoint was taken at 4 p.m.  $x'$ , with timestamp = 3 pm, is kept in main memory. This sibling, as modified, is written back to disk at 5 pm over the disk space of  $x$ .

If  $TS(x) < T_{chk} < TS(x')$ , i.e.,  $T_{chk}$  falls between the timestamps of two siblings, the sibling with the larger timestamp, i.e.,  $x'$ , is kept in main memory, and the other sibling, i.e.,  $x$ , is discarded from main memory. This case is illustrated in FIG. 6. The difference between FIG. 5 and FIG. 6 lies in the space on disk where the page is written back. The page is written over its own disk space in FIG. 6, i.e., over  $x'$ , because  $x'$  is not part of the last checkpoint state and can be written over, while  $x$  belongs to the checkpoint state and must be preserved. On the other hand, in FIG. 5,  $x'$  is written over the disk space of the other sibling, i.e., over  $x$ , for the reasons explained above. When the page is modified and written to disk, the timestamp relation is unchanged.

For a singleton page, if  $TS(s) < T_{chk}$ , i.e., its timestamp is smaller than  $T_{chk}$ , then it must be converted to a sibling form when it is modified, as shown in FIG. 7. This conversion is necessary because the checkpointed state must not be overwritten. After the conversion, the timestamp relationship becomes  $TS(x) < T_{chk} < TS(x')$ . When a singleton is converted to a sibling, it is copied to, say  $x$ , while its modified version is written over  $x'$ . On the other hand, if the timestamp of a singleton page is larger than  $T_{chk}$ , then its conversion to a sibling is not required because it is not a part of the checkpoint state.

Thus, each virtual page on disk is stored in either of the two different forms: sibling or singleton, depending on whether the virtual page is update-active or update-inactive, respectively. As these names suggest, virtual pages that have been recently updated are termed update-active, and those that have not been recently updated are termed update-inactive. To reduce the disk

space requirement, a sibling page is converted back to the singleton form when both siblings remain update-inactive for a long period, defined by a threshold parameter in the system. If  $TS(x) < TS(x') < T_{chk}$  and  $T_{chk} - TS(x') > (\text{threshold})$ , then the disk space manager reclaims their space by converting them into a singleton form. Singleton  $s$  contains the original contents and timestamp of the more recent sibling.

If there is a system crash, when the page table is restored it will point to the old disk location of the converted page. This old location may now contain unrelated information, as it may have been reallocated by the disk manager. Thus, it is necessary to keep a Forward-Page log (FP-log) outside of the virtual memory address space.

Whenever a sibling is converted to a singleton, and vice versa, an entry is written to the FP-log. This entry identifies the virtual page, and shows the old (before conversion) and new (after conversion) disk addresses. The FP-log allows the correct disk location of pages which have been converted to be recalculated after a system crash, even though the information in the page table portion of the checkpoint state is incorrect. Processing of the FP-log will be discussed further in connection with FIG. 8.

Whenever a sibling or singleton is modified, a sibling is written back to disk at a time after the last checkpoint. Since it is written after the last checkpoint but before a potential system crash, it might contain corrupted information. If there is indeed a system crash, that sibling must be expunged from the machine state by demoting its timestamp to  $-1$  during the post-crash recovery. When its timestamp is demoted to  $-1$ , the sibling is effectively rendered useless, and would be discarded if paged in during normal operation resumed after the post-crash recovery. If the timestamp were not demoted, the sibling would be retained in main memory, if paged in, and the roll back process would fail to recreate the checkpoint state.

Siblings whose timestamps must be demoted during the post-crash recovery can be located using a log, called the Expunge log. The Expunge log keeps track of a page which is written out to disk for the first time, if and only if, the page is part of the last checkpoint and the page has not yet been written to disk since the last checkpoint. This occurs if  $TS(x) < TS(x') < T_{chk}$  or  $TS(x') < TS(x) < T_{chk}$ . The virtual page identification and its disk address, along with the time of its writing to disk, are appended to the end of the Expunge log. The Expunge log must be maintained outside the virtual memory for it must survive system crashes. This can be done by writing it onto a different disk, for example, or to a specially set aside sector of the paging disk. In any event, the Expunge log is not addressable by the CPU.

If a system crash occurs before the next checkpoint, then as part of the post-crash recovery process, the Expunge log is processed: each page recorded in the log since the last checkpoint is located on disk, and the sibling with the higher timestamp is expunged by demoting its timestamp to  $-1$ . By demoting its timestamp to  $-1$ , the sibling is rendered useless.

Maintenance of the Expunge log requires disk accesses. This requirement can be amortized over a set of several pages. Instead of writing an entry to the Expunge log each time a page in the set is written out, a number of pages in the set can be queued up as they are on their way out to disk. Entries for all pages in the set can be first constructed and written into the Expunge

log, and then all the pages can be actually written out. Alternatively, a small non-volatile random-access memory can be used to buffer entries in the Expunge log. A background process would then be responsible for emptying this buffer to disk.

It is highly desirable to reduce the time required to complete the checkpoint operation. The major part of the checkpoint operation is the cleaning up of dirty pages by writing them to disk. Dirty pages are those which have been modified since the last checkpoint was taken. This time must be reduced so that there is no user-perceived pause attributed to the checkpointing process. Such pauses would be irritating for interactive users and disruptive for time-critical applications such as graphics and animation.

The time for checkpointing can be reduced by means of two techniques. The first is by making sure that the fraction of dirty pages in main memory at the checkpoint time is reasonably small by continuously cleaning up the dirty pages between successive checkpoints. This can be done by a continuously operating background process. A good heuristic to select a dirty page for cleaning up is "select a dirty page that has not been modified recently". The second is by cleaning up the dirty pages at the checkpointing time with no user-perceived pause, as explained below.

The checkpoint process may be initiated by an application or the system. At checkpoint time, a checkpoint process takes over. It saves all processor registers, the transient root, into a snapshot object, which is part of the virtual memory. The snapshot object resides at a fixed and known disk location, and can be optionally made a part of the transitive closure of the persistent root. All dirty pages in the main memory are then written to disk. Finally, for implementations using timestamps,  $T_{chk}$  is updated on disk to the current time, completing the checkpoint process. As described earlier,  $T_{chk}$  must either be successfully updated, or the old value retained. Immediately after the checkpoint operation,  $TS(x)$  and  $TS(x')$  must both be less than  $T_{chk}$  for all sibling pages, and  $TS(s)$  must be less than  $T_{chk}$  for all singleton pages. In the alternative embodiment, all sibling pages must be in either state Sib-1 or Sib-2. Optionally, all previously transient objects become persistent, as they are now members of the transitive closure of the persistent root by virtue of the transient root becoming a persistent object.

A further reduction of the perceived time used to take a checkpoint can be made by utilizing a copy-on-write technique to clean up dirty pages in main memory. Using this technique, all dirty pages in main memory are marked as copy-on-write; any attempt to write to them will cause a trap as explained below. This marking would generally be done in a page table, a data structure for keeping track of the location of virtual memory pages. The processor registers are then restored from the snapshot object and the normal operation resumes. All these operations involve only one disk operation to store the processor registers into the snapshot object on disk. Therefore, all these operations can be completed in well below 100 milliseconds on a typical winchester disk system, allowing the resumption of normal operation without any user-perceived pause.

The copy-on-write mechanism allows a very rapid resumption of the normal operation as no time is spent in writing the dirty pages back to disk before the normal operation resumes. After normal operation resumes, any attempt by the processor to modify the pages

marked as copy-on-write causes a trap. The trap handler first makes a copy of the page in main memory and this new copy is modified, but the original copy is not disturbed as it belongs to the checkpointed state. Any further attempts to modify are directed to this new copy. Until all pages marked as copy-on-write are cleaned up by writing them to disk, no newly created dirty pages are cleaned up. While the copy-on-write pages are being cleaned up, the normal operation continues. Cleaning up of dirty pages is preferably accomplished by a system routine running concurrently with user processes. Only when all pages marked as copy-on-write are cleaned up is the  $T_{chk}$  variable updated to be the current time, signaling the completion of a checkpointing operation. Note that at the end of a checkpointing operation, the entire machine state is stored on disk.

FIG. 8 indicates the activities during a post-crash recovery. After a system crash, diagnostics are run to detect any permanent hardware failures. If the diagnostics effort does not indicate a permanent hardware failure, or after any indicated failure is repaired, the FP-log is processed. This involves reading each entry in the FP-log, and using the new disk location contained therein to update the page table as stored in the checkpoint state. Next, each page indicated by the Expunge log is located on disk, and the sibling with the higher timestamp is expunged by demoting its timestamp to -1 as explained above. Singletons with timestamps greater than  $T_{chk}$  are also expunged. Then the state of the machine is rolled back to the state corresponding to the last checkpoint by restoring the processor registers from the snapshot object.

The system must be reconciled with the external world prior to restart. Part of the restored checkpoint state is related to the system configuration and I/O interfaces specific to the checkpoint time. This state must now be reconciled with the current time and configuration. For example, if the checkpoint state contains a timer variable, this variable must be adjusted to reflect the current time. At this point the normal operation from the checkpoint state can be resumed.

The problem of media or disk failures, as described earlier, can be addressed by archiving the last checkpointed state on disk to some other storage medium such as a streaming tape. This process is called the checkpoint archival process. If a disk failure occurs, a new disk can replace the damaged one, and the archived state can be used to initialize the new disk. Only those pages of the last checkpointed state that are dirty with respect to the last archived state need be copied to the archival medium. Pages that are dirty with respect to the last archived state indicate same with a flag in the page header. This flag can also be maintained in the page table. The next checkpoint cannot be taken until an ongoing checkpoint archival process, if any, is completed. The checkpoint archival process is a relatively infrequent operation and is expected to occur at most a few times a week, perhaps as an overnight operation, or when no user is working on the machine.

An alternative embodiment for selecting which sibling page to use from the disk works in a similar manner, but avoids the necessity for the use of page headers and the generation and maintenance of time stamps. The basis for operation of this mechanism is recognition of the fact that only four states are needed to describe the relationship between the timestamps of siblings  $x$  and  $x'$

and  $T_{chk}$ . The four states of sibling pages are: Both siblings written before  $T_{chk}$ , with  $x$  being older than  $x'$  in state Sib-1 and  $x'$  older than  $x$  in state Sib-2, as illustrated in FIG. 5;  $x$  written before the checkpoint and  $x'$  written after (state Sib-3) as shown in FIG. 6; and  $x$  written before the checkpoint with  $x$  written after (state Sib-4), which is the converse of FIG. 6. These four states are set forth in Table I.

TABLE I

State Name	3-bit Encoding	Description
bit position	2 1 0	
Sib-1	x 0 0	both written before checkpoint, $x'$ more recent
Sib-2	x 0 1	both written before checkpoint, $x$ more recent
Sib-3	0 1 0	$x$ written before checkpoint $x'$ written after checkpoint
Sib-4	0 1 1	$x'$ written before checkpoint $x$ written after checkpoint
Sing-1	1 1 0	$s$ written before checkpoint
Sing-2	1 1 1	$s$ written after checkpoint

The actual value of the timestamps used in the previous embodiment is important only for the purposes of comparison with timestamps of siblings and the most recent checkpoint. The alternative embodiment simply abstracts the relationships between the timestamps into four states as described above. Only two bits are necessary to fully distinguish among the four states of a sibling page. Table I also shows the two states which are necessary if singleton pages are to be used. This requires that the states be encoded using three bits. This is much smaller than the size of the timestamps associated with each page in main memory, and provides numerous advantages as will be shown below.

As in standard virtual memory management mechanisms, a page table is used to locate pages on disk when a page fault occurs. The page table is a data structure which associates virtual page addresses with disk addresses, and typically contains other status information. A simplified form of such a page table data structure is shown in FIG. 10. When a page fault occurs, the virtual page address is located in the page table, typically through hashing or other search techniques, and the associated physical disk address is retrieved. In the preferred embodiment, the bits necessary to describe the state condition of the siblings is contained in three state bits in the page table. The physical disk address information is always directed to one sibling, for example  $x$ , and the other sibling,  $x'$ , is always located at some disk position which is calculable from the physical disk address of  $x$ . For example, both siblings can be stored in adjacent blocks, or the same track/block for different disk heads, or in the same position on different disk drives. As an alternative, both physical disk addresses can be stored in the page table, so that sibling pages may be scattered randomly over the disk. This alternative method has the drawback of increasing the size of the page table.

When a page fault occurs, instead of reading in both sibling pages and discarding one, as described in connection with the first embodiment, only the correct disk page is read in. The correct disk page address is calculated by examining the state bits, and choosing the correct sibling. The sibling which is read into memory, and which is written to for a page out operation, is the same as that described in the timestamp embodiment: only the

method of calculation for choosing the appropriate pages modified.

When pages are written to disk, the relative time-or-write of the sibling pages and  $T_{chk}$  may change, in the manner illustrated in FIGS. 5 and 6. This means that the state bits will need to be modified during a page out operation. Referring to Table I, if state is Sib-1, both siblings were written before the last checkpoint, with  $x'$  more recent. Thus,  $x'$  is fetched into main memory. If  $x'$  is modified, it is written out to the disk space occupied by sibling  $x$ , with the state being changed to Sib-4. This preserves  $x'$ , which is part of the checkpointed state.

An analogous situation occurs when a page is read in and state is Sib-2. When state is Sib-2, and the dirty page is written to disk, the state of the page changes to Sib-3.

If a page is read in, and the state is Sib-3 or Sib-4, the state of the page remains unchanged when a dirty page is written out to disk. It will be recognized that these state changes correspond to preservation of the checkpointed state as described for the embodiment utilizing timestamps.

Table I indicates states which are used for singleton pages (Sing-1 and Sing-2). In many cases, a system will use only sibling states, which would require only two state bits in the page table. If singleton pages, as described with the previous embodiment, are used, a third bit must be added and stored in the page table. The two possible states for a singleton page are:  $s$  written before or after the most recent checkpoint. These correspond to states Sing-1 and Sing-2 in Table I.

If a page fault indicates a page is to be read in having state Sing-2, that page can simply be read into memory. It is not necessary to make a copy and convert the page into sibling form. When the page is written out again, the state remains Sing-2, and the page is written into its old disk location.

If a page having a state Sing-1 is paged in, it is necessary to convert that page to sibling form. This is because the singleton page is part of the checkpointed state, and must be preserved. The conversion operation is the same as shown in FIG. 7. In that example, the state is changed from Sing-1 to Sib-4 as shown in Table I.

When a checkpoint is taken,  $T_{chk}$  will be more recent than all pages which have been written to disk. Therefore, the states of some pages will need to be changed as part of the checkpoint operation. Once the checkpoint is completed, all pages will be in either Sib-1, Sib-2 or Sing-1 states. Siblings that were in state Sib-3 changed to Sib-1, and those who were in Sib-4 changed to Sib-2. Pages which were originally in Sib-1, Sib-2, or Sing-1, do not change state.

If a page has been in state Sib-1 or Sib-2 for a long time, the older sibling contains information which has been obsolete for a long time. Such a page can be converted to the singleton form for the reasons described in connection with the embodiment using timestamps. The contents of the more recent sibling are copied to a singleton page,  $S$ , which will be in state Sing-1.

The bits marked "x" can be used for deciding when a conversion to the singleton form should take place. A procedure, which could be a background process, constantly scans all entries in the page table with a preselected frequency. This frequency is preferably quite low, and can be set as desired according to the system designed. This background process inspects the "x", or "idle bit", for all pages which are in the states Sib-1 or Sib-2. If the idle bit is "0", it is changed to "1" and the scan continues. If the idle bit is found to be a "1", that

page has not been updated since the previous scan through the page table, and that page is a candidate to be converted to singleton form. When a singleton page is converted to sibling form, or pages in states Sib-3 or Sib-4 are changed to the corresponding state Sib-1 or Sib-2 as result of a checkpoint being taken. The idle bit is set to zero. The length of time which an unmodified page will remain in sibling form before being converted into singleton form depends entirely on the scan rate of the scanning process.

The advantages of the second embodiment over that of the first are several. It is not necessary to maintain a timestamp for each disk page, so that no page header is needed. Since there is no page header, there is no requirement that the disk controller be able to perform a scatter-gather operation as described earlier. Since only one sibling is read in, as opposed to reading two and then discarding one, it is not necessary that siblings be physically contiguous on the disk. They may be separated, and it is only necessary that the location of one sibling can be calculated from that of the other. Alternatively, the page table can be enlarged slightly to include both siblings in the page table. Since the simple case requires only two extra bits in the page table, or three if singleton pages are used, and a few extra bits are available on many systems, the second embodiment can be implemented on any existing virtual memory system.

Generally, pages in sibling form and singleton form will be stored in different locations on disk in order to optimize disk usage. A page table within the virtual memory address space indicates the disk location of each allocated virtual page. The page table is a persistent object, and is checkpointed as are all other objects. When a sibling is converted to a singleton, or vice versa, the page table is updated to point to the new disk location for that page, and the old location is deallocated and made available for later use.

When the embodiment using state bits in the Page Table is implemented, it is not necessary to use the Expunge log. Pages written after the checkpoint in states Sib-3, Sib-4, and Sing-2 may be corrupted and should be ignored. Simply rolling the page table back to the checkpointed state restores the correct state as defined in Table I. No changing of timestamps or states is necessary.

#### TECHNICAL ADVANTAGES

A virtual memory using a state indicator in a page table can be made recoverable from system crashes. This can be done without placing any additional burden, such as a scatter-gather capability, on the disk controller. This allows for relatively easy implementation of a persistent memory on nearly any existing virtual memory system. Some advantages of this system over one using timestamps include: no timestamps or page headers are necessary (which uses less memory on disk and in main memory), no expunge log is needed, and the overall implementation is relatively straightforward.

Although a preferred embodiment has been described in detail, it is understood that various substitutions, additions and modification may become apparent to those skilled in the art. These changes may be made without departing from the spirit and scope of the invention, which is defined by the appended claims.

What is claimed is:

1. A computer system having a recoverable virtual memory, comprising:



- (a) a virtual memory having a random access memory and a backing store;
  - (b) a plurality of pages stored in the backing store, each of said pages corresponding to a virtual memory page;
  - (c) a plurality of memory objects allocated within said virtual memory, said memory objects interconnected by pointers;
  - (d) a persistent root having a pointer to a memory object, wherein all memory objects in the transitive closure of said persistent root define a set of long term objects not subject to reclamation by the system during garbage collection;
  - (e) means for taking checkpoints of the current state of said virtual memory at selected times, said means including means for materializing a sibling page corresponding to each of a plurality of said virtual memory pages, wherein the latest written version of a virtual memory page and its respective sibling page which was written before a current checkpoint state is used to define the checkpoint state; and
  - (f) a state indicator identified with each virtual memory page and each sibling page for indicating which of said virtual memory or sibling pages stored in the backing store comprise a most recent checkpoint.
2. The computer system of claim 1, wherein the computer system is reset to the most recent checkpoint state after a system failure.
3. The computer system of claim 1, wherein the backing store is a paging disk.
4. The computer system of claim 1, further comprising a page table for indicating the correspondence between virtual memory pages and backing store pages, wherein said state indicators are located in said page table.
5. The computer system of claim 4, wherein said state indicators comprise a plurality of  $n$  bits, whereby  $2^n$  states can be defined by said state indicator.
6. The computer system of claim 1, wherein sibling pages share a single state indicator with their associated disk pages.

7. The system of claim 1, wherein said persistent root is a memory object located within the virtual memory space.
8. A method for recovering from a system failure in a virtual memory having a plurality of virtual pages stored on a backup store, comprising the steps of:
- (a) providing a persistent root having a pointer to an object in the virtual memory, wherein objects in the transitive closure of the persistent root define a set of nontransient objects which are not subject to reclamation by the system during garbage collection;
  - (b) identifying a state indicator with each virtual page when that page is written to the backup store;
  - (c) taking periodic checkpoints of the virtual memory by storing the states of any registers and the persistent root onto the backup store at a predetermined location, and storing state indicators for the virtual pages at a predetermined location;
  - (d) saving a current version of each virtual page in existence when the periodic checkpoint is taken; and
  - (e) when a system failure occurs, restoring a checkpoint state of the virtual memory by discarding all virtual pages whose state indicator indicates that they were written to the backup store after the most recent checkpoint was taken.
9. The computer system of claim 8, wherein step (c) includes the step of changing the state indicator for virtual pages containing the page table to indicate that it is a part of the checkpoint state.
10. The computer system of claim 8, wherein step (d) comprises the steps of:
- (d.1) materializing sibling pages corresponding to each of said virtual pages which do not have respective sibling pages, and writing said sibling pages to the backing store when such virtual pages are updated, wherein the sibling pages contain updated versions of the virtual pages and the virtual pages are unchanged; and
  - (d.2) for virtual pages already having respective sibling pages, leaving unchanged the virtual page or the sibling page which was most recently stored before the last checkpoint was taken, and storing an updated version of the virtual page in the remaining of the virtual page or sibling page.
- \* \* \* \*